PES
UNIVERSITY

Dissertation on

“Title of the project”

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE18CS390A – Capstone Project Phase - 1

Submitted by:

Name 1	**<SRN 1>**
Name 2	**<SRN 2>**
Name 3	**<SRN 3>**
Name 4	**<SRN 4>**

Under the guidance of

Prof. Guide Name
Designation
PES University

January - May 2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

‘Title of the project’

is a bonafide work carried out by

Name 1	**<SRN 1>**
Name 2	**<SRN 2>**
Name 3	**<SRN 3>**
Name 4	**<SRN 4>**

in partial fulfilment for the completion of seventh semester Capstone Project Phase - 1 (UE18CS390A) in the Program of Study - Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2021 – May. 2021. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6th semester academic requirements in respect of project work.

Signature	Signature	Signature
<Name of the Guide>	Dr. Shylaja S S	Dr. B K Keshavan
Designation	Chairperson	Dean of Faculty

External Viva

Name of the Examiners **Signature with Date**

1. ____________________________ ____________________________

2. ____________________________ ____________________________

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"Title of the project"** has been carried out by us under the guidance of <Prof. Guide Name, Designation> and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2021. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

<SRN 1>	**<Name 1>**	**<Signature>**
<SRN 2>	**<Name 2>**	**<Signature>**
<SRN 3>	**<Name 3>**	**<Signature>**
<SRN 4>	**<Name 4>**	**<Signature>**

ACKNOWLEDGEMENT

I would like to express my gratitude to Prof. <Guide Name>, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE18CS390A - Capstone Project Phase – 1.

I am grateful to the project coordinators, Prof. <Project Coordinators Name>, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Shylaja S S, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

With the advent of the sphere of bioinformatics, of there is a new confluence engineering and biology that has led to remarkable changes in the way researches approach difficult and time consuming problems. The pharmaceutical industry has highly benefitted from this. Research and opened new gates with regards to how we look at drugs and their applications. With this window of opportunity, comes a big challenge: information on these drugs are not computationally friendly. In this work we take advantage of the structural and functional aspects of a drug to generate a drug embedding, an accurate representation which will serve as a gateway to other bio informatic applications like drug discovery, drug target interactions, drug reprofiling and drug repositioning. We employ machine learning techniques to learn the embeddings of the drugs and validate their efficacy by testing against a known biological classification. Additionally, we are aiming to solve an important bioinformatic application which is: early and accurate identification of potential adverse drug reactions (ADRs) for combined medication which is vital for public health. Since most clinical trials focus on a single drug and its therapeutic effects, most drug-drug interaction induced ADRs go unnoticed until the drugs are actually approved. This has been one of the top 10 reasons on death in United States according to the studies. To solve the same problem we employ various machine learning models to predict drug-drug induced ADRs using the structural and functional characteristics of drugs.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The field of bioinformatics has been revolutionizing the way we look at drugs and their use cases. Bioinformatics is an interdisciplinary field at the confluence of biology and engineering that develops computational methods for analysing biological data. This emerging field provides useful information for the analysis of the molecular basis of diseases, the search for target proteins, and the analysis of the effect of drug therapies. It has also been instrumental in the advance of disease diagnosis and prognosis as well as therapy selection. With the wide range of life altering use cases the field presents to the world, it has become indispensable in today's world.

There has been a surge of computational breakthroughs to support the ever-evolving field over the years thanks to the current advancements that make it possible to look at drugs from all aspects-structural and functional.

Bioinformatic applications like drug discovery, drug target interactions, drug reprofiling and drug repositioning call for years of research and huge monetary investments from pharmacological companies. In the past decade there has been an enormous increase of financial investments in pharmaceutical R&D. This increase in investment has not reflected in the number of newly approved drugs. Drug approval process is very tedious and picky. This is the motive behind using computation to aide these conventional wet lab research is to increase the number of drug approvals or find new uses for already approved drugs, which will save a lot of time and monetary expense.

In the cut-throat commercial drug industry, designing chemicals with desired characteristics is a bottleneck in the development of new drugs. Drug design is still driven by intuition, chance and years of research. However, processing of structural and functional descriptors of drugs can lead to at an expedited development. This requires good 'Representation Learning' which is a precursor to help these descriptors become computationally friendly and meaningful.

Existing structural and functional indicators of drugs are not in consumable formats for mathematical processing. They are categorical, mostly through manual process. In this work, we explore the idea of drug embedding, a representation of a drug. The performance of machine learning methods relies heavily on the input data. Therefore, 'Representation Learning'- the design of data pre-processing and data transformation is of great concern to ensure that the data representation can support machine learning algorithms. This is one of the biggest challenge of bioinformatics and with our work we attempt to try to find a solution to this.

An important application of bioinformatics is classifying drugs into their respective adverse effects. Adverse drug reactions (ADRs) are unintended and harmful reactions caused by normal uses of drugs. ADRs caused by individual drugs and drug combinations constitute one of the top 10 causes of death in the United States. With an increase in the number of drugs used in pharmacology, there has been an increase in the number of ADRs due to drug-drug interactions. The issue with drug-drug induces ADRs is that they are not observed in the clinical trials which usually focus on a single drug and its therapeutic effect against a particular disease or condition

Clinical trials are conducted for each drug before approval of the drug which is conducted in 3 phases as described: Phase I clinical trials investigate the safety profile of a candidate drug on a small group of volunteers. Phase II trials evaluate its safety and efficacy in a larger group of volunteers. Finally, in Phase III clinical trials, these effects and ADRs are monitored in a large group of selected volunteers. Drugs that successfully pass these hurdles can then be approved for general clinical use. In fact, although most ADRs induced by individual drugs have been discovered and carefully monitored in clinical trials before drug approval, information on nearly all ADRs induced by drug-drug interactions (DDIs) has been generated after drug approval. This poses continuous and serious risks to patients' health.

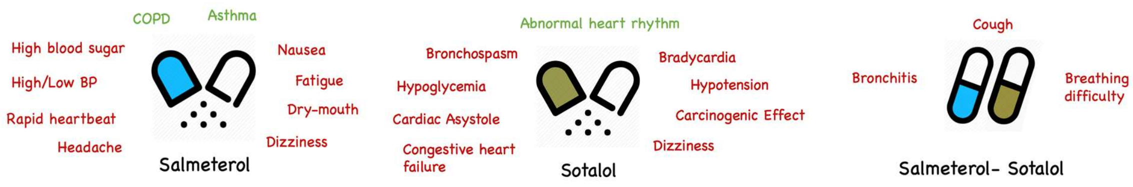


Figure 1: ADRs caused by drugs when used separately as compared to used together

Furthermore, in most scenarios the ADRs observed by individual drugs completely differs from the ADRs observed when both the drugs are taken together. Figure 1 describes one such example where the drug Salmeterol and Sotalol cause totally different ADRs when taken together as compared to when taken individually.

CHAPTER 2

PROBLEM STATEMENT

Wet lab testing has always been a costly and time intensive affair. Computation has attempted to bridge the gap between wet lab testing and timely results. Computational techniques can nudge researchers in the right direction without costing them a fortune of their time and money. One such example is drug repurposing. There exists millions of chemical compounds each with a specific set of properties that give it an appropriate use case. Finding drugs correctly suited for different requirements is generally considered a job for chemical laboratories where compounds are tested for various properties. It is difficult to accomplish this without extensive testing because drugs cannot be considered similar despite superficial similarities such as common functional groups. Two drugs with similar structures may not be suitable for the same application.

A solution for determining similarities between drugs stems from examining different legs of the drug like its structure and gene expressions. The structure of a drug can be documented by a simple string of atoms connected with bonds. A gene expression is obtained by applying drugs to RNA strands and observing proteins released. Two drugs that can be used for similar applications may have some inherently common patterns within these structural strings or gene expressions.

Since these inputs are quite varied and are not computationally friendly a pre cursor to such heavy bioinformatics problem would be to find a suitable multi-purpose representation of a drug. This embedding would serve as a gateway to determine different drug-target interactions, similarities between drugs and bases for drug discovery.

The problem that we are hoping to solve through our research for this project is to try to extract inherent patterns from structural and functional information on a drug using AI. Deep learning architectures will be employed to embed this information such that the embeddings of different drugs will be able to partake in many bioinformatic applications.

We analyse different types of information available on drugs and their representation, understand the nuances of the data to glean the most suitable dataset for our problem and compare different deep learning methods to arrive at an architecture that will give an embedding which will capture the details of the drug and can be scaled to different applications.

Most studies of ADR prediction are based on using drug-attributes based on chemical features of drugs like use of ECFP on organ-based ADR prediction, use of CACTVS and SCCA or drug-attributes based on biological features of attributes like Drugs with similar ADR profiles tend to have similar protein target.

On the other hand, work on ADR caused by combined medication, which occurs when individually safe drugs interact pharmacokinetically or pharmacodynamically, is limited to data mining and uses no X-omic study as per our literature study. Pharmacologic databases provide rich useful resources for identifying ADRs of combined medication, owing to the openness, high data quality, and coverage for both novel and rare drugs.

In this study we used machine learning models with the transcriptomic data from L1000 to predict the risk of ADR on combined medication of two drugs. We also performed a comparison of the predictive performances of five state-of-the-art and perhaps most commonly employed ML algorithms, including Decision Trees, Random Forests, Naïve Bayes, Logistic Regression as well as Stochastic Gradient Descent.

CHAPTER 3

LITERATURE SURVEY

In this chapter, we present the current knowledge of the area and review substantial findings that help shape, inform and reform our study.

3.1 Background on Embeddings and Drug Repurposing

This section details the papers read to gain information on background, data used, and the current methodologies being used in the field of bioinformatics.

3.1.1 Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery [1]

Zheng Xu et. Al

In this paper, the authors propose a novel unsupervised molecular embedding method (unsupervised seq2seq fingerprint method), to simply translate a SMILE string into itself. The intermediate vector generated can be considered as a continuous feature vector for each molecule of a drug to perform classification.

Architecture of the network: GRU cell is used, instead of LSTM, to accelerate the training process. Attention Mechanism is employed to centralize the finger-print space. Dropout layer is added to overcome the over-fitting.

The benefits of the seq2seq fingerprint are three folds:

1. The training phase of seq2seq fingerprint is unsupervised: completely label-free.
2. It is data-driven, eliminating the reliance on expert’s subjective knowledge.
3. Unlabelled data is unlimited, the deep learning network can be trained well.

The author’s network takes a very long time to train. The length of the embeddings are in the set of (512, 784, 1024) which is quite large and requires more computation and robust hardware.

3.1.2 Predicting New Molecular Targets for Known Drugs [2]

Michael J. Keiser et. al

The authors used a statistics-based chemo-informatics approach to anticipate associations between drug and targets. Most drugs had no significant similarities to most ligand sets. However, 6,928 pairs of drugs pairs of drugs and ligand sets were similar, with expectation values (E-values) better than 1×10^-10.

Not all the new off-targets predicted by their research are unanticipated. A third of their drugs which were predicted active on their off-targets are false positives when verified with wet lab tests.

3.1.3 Drug Target Identification Using Side-Effect Similarity [3]

Monica Campillos et. Al

The authors explore the relationship between drugs and their protein targets. It is observed through this research that there is an inverse correlation between side-effect frequency and the likelihood of two drugs to sharing a protein target. The authors tested the predictive power of this side-effect similarity measure on their reference set of 502 drugs with known human targets and observed a clear correlation between side-effect similarity and the likelihood that two drugs share a protein target. The authors observed in their reference set that chemically similar drugs [according to the two-dimensional (2D) Tanimoto chemical similarity score] are likely to have the same targets.

This study presented high accuracies (88% AUC) on the enzyme dataset however, the similarity metric was based off of only 502 drugs.

3.2 Generating Embeddings for SMILES

This section details the survey of the current state-of-the-art methods for various bioinformatic applications that employ SMILEs as their primary dataset.

3.2.1 SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties [4]

Garrett B. Goh et. Al

Based off of NLP techniques which gather grammar related features not explicitly but through training a network, so does Smiles2Vec train an RNN model to learn and understand the grammar in SMILEs. Through training, the network is expected to learn features that could be useful in understanding the chemical properties of the molecule. Embeddings are tested for properties like solubility, activity, toxicity and solvation energy.

The results presented by this paper show embeddings with a good representation of the molecule. Prediction of functional properties have not been attempted by this paper and could be a possible enhancement. Other feature gathering models like CNN could be utilised to better the quality of the embeddings.

3.2.2 SPVec: A Word2vec-Inspired Feature Drug-Target Interaction Prediction [5]

Yu-Fang Zhang et. Al

SPVec uses the word2vec model to embed smiles into embeddings that represent the molecule. The skip gram model coupled with a negative-sampling has been used to achieve a robust model that creates representative smiles. The learned embeddings are tested for their ability to predict various chemical properties using algorithms like random forest, gradient boosted decision trees and DNNs. Promising results have been presented.

While the results demonstrated in this paper showcase a decent ability of the embeddings to predict chemical properties, the functional information held by the embedding has not been tested and could be a possible future enhancement.

3.2.3 Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition [6]

Sabrina Jaeger et. Al

Employing the core strategies in natural language processing, the paper describes mol2vec a methodology for converting a molecule into a representative vector. Using unsupervised machine learning techniques, each substructure in the molecule is encoded into a vector using the skipgram model. Averaging the substructure embeddings gives a molecule level embedding which has been used to represent the molecule and it's efficacy has been verified using the embedding's ability to predict it's properties.

The representative ability of embeddings of substructures could be further observed by changing the radius at which substructures are obtained. Novel embedding techniques other than skip gram could be tested for better performance.

3.2.4 SWeeP: Representing Large Biological Sequences Datasets in Compact Vectors [7]

De Pierri CR, et. al

The paper looks towards representing various biological sequences like gene sequences in a compact form while retaining all or most of the information held in the original embeddings. The embedding is created by locating indices in the sequence to make a high dimensional vector which is later projected to a lower dimensional space to obtain a compact vector.

Spaced word projections works well but it's purely mathematical approach could be combined with machine learning techniques to further enhance the quality of the embeddings obtained.

3.3 Generating Embeddings for Gene Expression

This section details the survey of the current state-of-the-art methods for various bioinformatic applications that employ transcriptomic data as their primary dataset.

3.3.1 Drug Repurposing Using Deep Embeddings of Gene Expression Profiles [8]

Yoni Donner et. Al

Here, the authors report a new method for measuring functional similarity between drugs based on gene expression data using deep neural networks to learn an embedding that substantially denoises expression data, making replicates of the same compound more similar.

The method uses unsupervised deep learning method with each layer being a densely connected with SELU activation. Once the embeddings were retrieved, their method could identify drugs with shared therapeutic and biological targets even when the compounds were structurally dissimilar, thereby revealing previously unreported functional relationships between compounds. Their method presented good results, 95% accuracy for ATC classes.

Biologically, they did not account for the fact that a drug acting on a particular cell line would be comparable to the same drug acting on another cell line.

3.3.2 A Large-Scale Gene Expression Intensity-Based Similarity Metric for Drug Repositioning [9]

Chen-Tsung Huang et. al

Biological systems like gene expression data are usually robust to perturbations and the current similarity techniques that are in use, don't capture this. The authors propose that intensity-based similarity metric surpasses other standard metrics in drug clustering. This metric was applied to compare thousands of compounds for drug repurposing. The new metric emphasizes the genes exhibiting the greatest changes in expression in response to perturbation.

3.3.3 Gene2vec: Distributed Representation of Genes Based on Co-Expression [10]

Jingcheng Du et. al

The authors proposed a method that utilizes gene co-expression to generate a distributed representation of genes. The distributed representation of genes could be useful for more bioinformatics applications. Inspired by the success of word embedding, they intend to produce an embedding of genes. Since genes don't have an equivalent of a 'sentence', they use the notion of co-expression. This is analogy of the concept of co-occurrence in natural languages.

The authors were successful in creating an embedding, however for the purposes of training gene co expression can prove to be a very large matrix which is not easily managed.

3.3.4 Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data [11]

Alexander Aliper et. al

In this paper the authors demonstrate classification of transcriptomic data into therapeutic categories. They used the perturbation samples of 678 drugs across A549, MCF-7, and PC-3 cell lines from the LINCS Project and linked those to 12 therapeutic use categories derived from MeSH. They were able to prove that gene expression data needs a deep learning network to classify accurately. A machine learning model like SVM will not suffice.

Given the skewed nature of the MeSH dataset as well as the small number of drugs (678) used to train the network. It may not scale well.

3.3.5 Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics [12]

Ehsaneddin Asgari et. al

In this method, the authors present a representational learning method to generate continuous vector representation for drugs using an artificial neural network. To evaluate this method, they apply it in classification of 324,018 protein sequences obtained from Swiss-Prot belonging to 7,027 protein families, where an average family classification accuracy of 93%±0.06% is obtained, outperforming existing family classification methods. The resulting embeddings prove to cluster protein sequences with well-defined boundaries.

3.3.6 Drug Repositioning: A Machine-Learning Approach Through Data Integration [13]

Francesco Napolitano et. al

The noisiness and scarcity of the gene expression are important limitations to using solely gene expression in computation. The authors propose a novel computational approach to predict drug repositioning based on integrating multiple layers of information:

i) Distances of the drugs based on how similar their chemical structures are.
ii) Closeness of their targets within the protein-protein interaction network
iii) Correlation of the gene expression after treatment.

Their classifier reaches high accuracy levels (78%), but this method may not be extended to other classification systems as the target closeness information for this particular dataset can be obtained from protein-protein interaction network. There may not be similar networks that can provide information about other targets.

3.4 Adverse Drug Reactions

This section details the survey of the current state-of-the-art methods for bioinformatic application of classifying drugs into their adverse effects.

3.4.1 Systematic Drug Repositioning Based on Clinical Side- Effects [14]

Lun Yang, Pankaj Agarwal

Clinical side-effects (SEs) provide a human phenotypic profile for the drug, and this profile can suggest additional disease indications. The authors extracted 3,175 SE-disease relationships by combining the SE-drug relationships from drug labels and the drug-disease relationships from PharmGKB.

The AUC was above 0.8 in 92% of these models. The method was extended to predict indications for clinical compounds, 36% of the models achieved AUC above 0.7. The authors built Naive Bayes models to predict indications for 145 diseases using the SEs as features. The AUC was above 0.8 in 92% of these models.

3.4.2 Predicting ADR of Combined Medication from Heterogeneous Pharmacologic Databases [15]

Zheng Y et. al

For the task of predicting ADRs more effectively, the authors used highly credible negative samples (HCNS-ADR), fused heterogeneous information from various databases and represent each drug as a multi-dimensional vector according to its chemical substructures, target proteins, substituents, and related pathways first. Then, a drug-pair vector is obtained by appending the vector of one drug to the other.

Next, they construct a drug-disease-gene network and devise a scoring method to measure the interaction probability of every drug pair via network analysis. Drug pairs with lower interaction probability are preferentially selected as negative samples. They did PCA for negative and positive samples.

Finally, a classifier is built for each ADR using its positive and negative samples with reduced dimensions. The models showed significant improvement with HCN-ADR samples.

The authors were able to get higher performance using the HCN-ADR samples created.

3.4.3 Predicting Adverse Drug Reactions Through Interpretable Deep Learning Framework [16]

Sanjoy Dey et. al

I In general, the basic steps of ADR prediction based on structural information can be broken down into two stages. First, each drug molecule is represented in a suitable feature vector based on its chemical structure. Second, a machine learning algorithm is applied on the resulting feature space to predict ADRs. Since, there is not a lot of work done regarding the first part of the project in this paper, they have developed machine learning models including a deep learning framework which can simultaneously predict ADRs and identify the molecular substructures associated with those ADRs without defining the substructures a-priori.

They used CNN to represent the complex structures in fixed length vectors and simultaneously use deep learning to analyse the fingerprints and relate them to ADR. In particular, they design R hidden layers in the deep learning framework, each corresponding to a particular radius. Therefore, their framework can search for all possible substructures up to radius R by successive increment of the radius of the substructure by one in each layer of neural network. Afterward, the similar structures are summarized into a final feature representation called fingerprint.

At each step (radius), they use an additional attention mechanism step to map the contribution of each of the substructures into the final fingerprint. Finally, the fingerprints are assessed in terms of how well they can predict ADRs. They built a predictive model using L2-norm regularized logistic regression method for each ADR separately using those fingerprints as features.

3.4.4 Detecting Potential Adverse Drug Reactions Using a Deep Neural Network Model [17]

Chi-Shiang Wang et. al

The objective of this study was to identify a method to detect potential ADRs of drugs automatically using a deep neural network (DNN). They demonstrated the usefulness of the proposed representation by inferring two types of relations: a drug causes a side effect and a drug treats an indication. To predict these relations and assess their effectiveness, they applied 2 modelling approaches: multi-task modelling using neural networks and single-task modelling based on gradient boosting machines and logistic regression. They used MeSH and SIDER dataset.

The drug representation includes co-occurrence frequencies between each drug and all other MeSH terms. These numbers have to be normalized to account for the variability in the total number of drug occurrences. They implemented 3 normalization methods:37 maximum term-frequency normalization (Max-TF), in which each coordinate in the representation vector is divided by the maximum value of that vector; Log + Max-TF, in which the logarithm of the terms count is taken followed by Max-TF; and TF–inverse document frequency, which is commonly used in text mining and information retrieval for term normalization.

To examine the prediction performance of the drug representation on the 2 tasks, they used 3 types of models: Multilayer fully connected NN architecture, GBM, implemented using the LightGBM package.

CHAPTER 4

DATA

This chapter serves to describe the data under consideration. Understanding the way all the data work is vital in the process of creating a good solution to the problem at hand.

4.1 Overview

A drug is characterized by its structure, functional abilities and the side effects it causes. In order to generate a good embedding of a drug, we explore all of the above and keep what proves to be truly representative of a drug and what is not mathematically in concord with our problem.

4.2 SMILES: Structural Indication of a Drug

The simplified molecular-input line-entry system (SMILES) is a specification in the form of a line notation for describing the structure of chemical species using short ASCII strings. Encoded in each SMILES string is structural information that can be used to predict complex chemical properties.

SMILEs are representations of the structure of a drug. It is a "chemical language" that encodes structural information of a chemical into a compact text representation that is easy to consume. SMILEs use a strict grammar to maintain consistency. The alphabets (example, C, O, H) denote atoms, and in some cases also what type of atoms. Special characters (like '=') denote the type of bonds (single, double, triple). Round brackets encapsulating numbers, and side chains denote rings. From this structural information, more complex properties can be predicted.

Example- Aspirin: CC(=O)OC1=CC=CC=C1C(=O)O

The SMILES data is extremely useful for mathematical representation of drugs as it contains both spatial and sequential data. We can leverage this to our advantage.

Our dataset holds 1004 smiles with their respective ATC classification (refer Section 4.5.1)
This is the main dataset that we will be concerning with.

4.3 LINCS: Functional Indication of a Drug

The dataset under consideration to represent the functional aspect of a drug, i.e. how it interacts with cell lines is procured from the NIH Library of Integrated Network-Based Cellular Signatures Program. The L1000 assay (which is one of the many gene inclined datasets available as part of the NIH Library) measures mRNA transcript abundance of 978 “landmark” genes from human cells. Measurements of these 978 "landmark genes” are applied to an inference algorithm to infer the expression of 11,350 additional genes in the transcriptome. 1.6 million profile expressions are obtained by measuring mRNA abundance in perturbed cells.

Table 1: Statistics of the L100 Assay

Number of Perturbagens[1]	2107
Number of samples per perturbagen on an average	42
Dimension of each sample	11,350
Total number of samples	118,500

The dataset is divided into levels with different levels representing different aspects.

Level 1: There are various Luminex graphs, where x axis represents time and the y axis represents fluorescence, generated for different profiles.

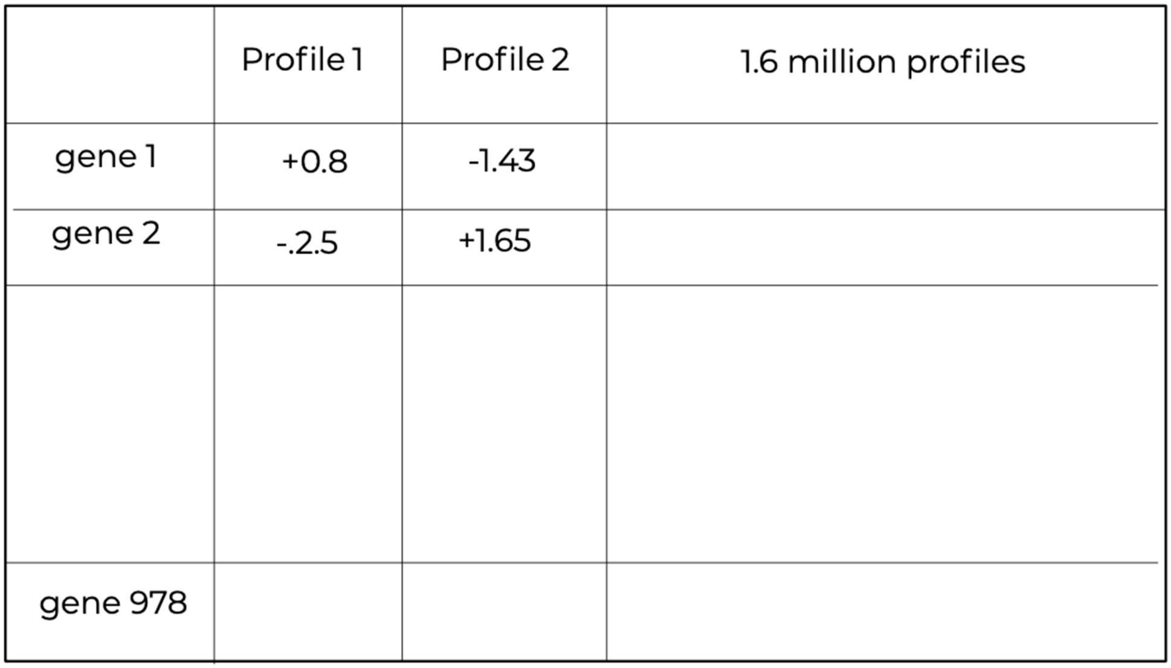
[1] A perturbagen is a drug compound

Level 2: Reduced the 22k genes to 978 landmark genes for ease. Convert the above graphical data into tabular data [978 x 1.3 million]. 1.3 million columns are the various profiles. Control genes are those sequences that have been determined as unchanging on addition of any perturbagen in any of the genes. There are 80 such control genes which are represented as 80 columns in the dataset.

Level 3: This data has inferred information for 12K genes and 1.3 million profiles form level 2 data through a series of steps involving standardisation, normalisation and in-Ferring the remaining 11k genes through a multiplication with a weighted matrix.

Level 4: Indicates the up and down regulation in the gene expression on adding perturbagen. After getting the standard gene expression values from the control genes by average all the control probe values for all genes and average control probes for each gene, we perform a Z score like operation to give us the regulated value for the table.

Level 5: The replicates present in the level 4 data are consolidated which reduces the data dimension. For the purpose of this problem, Level 5 data is considered.



	Profile 1	Profile 2	1.6 million profiles
gene 1	+0.8	-1.43	
gene 2	-.2.5	+1.65	
gene 978			

Figure 2: Representation of the L1000 Assay

4.4 Combined Feature Set for ADR Detection

Collection of **transcriptomic data** and ADR-drug association Gene expression profiles for 978 landmark genes were extracted from **NIH funded LINCS L1000** dataset for level 5. The level 5 data provide transcriptomic data for 978 landmark genes for 20,413 small molecules in the form on consensus signature. A signature is a continuous values for the gene expression extracted from before and after the action of small molecule. Gene expression for a perturbagen is the expression for the strongest signature for perturbagen. The strongest signature for perturbagen is defined as the signature with highest "distil_ss" value for the signature common to given perturbagen. The feature "distil_ss" gives the magnitude of the difference of the differential expression for the signature from the average of DMSO treated control sample. The highest magnitude corresponds to most effective perturbagen and can be used irrespective of the cell line, concentration and duration of treatment. The preparation of gene expression data has been inspired by Wang et. al.

Two-dimensional **representation of the chemical structure** of the perturbagens were extracted in the form of **Simplified molecular-input line-entry system (SMILES)** from L1000 metadata for phase one. SMILES, which is present as metadata in L1000 dataset for small molecules, were found and converted to molecular fingerprint, namely **Molecular ACCess System structural keys (MACCS) - 166-bit structural key,** with each bit associated with SMARTS pattern. This conversion was made using the **Open Babel package.**

The genes from the LINCS dataset were represented using **Gene descriptors, specifically Gene Ontology**, using **Principal Angle Enrichment Analysis (PAEA)** method. Gene Ontology as gene descriptors were used in many publications for representations.

The feature set for the prediction consisted of the combined set of chemical fingerprints in the form of bit array of MACCS, molecular descriptors in the form of Gene Ontology and the effect of the drugs on genes in the form of RNA-expression from L1000 project.

From the feature set, we get transcriptomic data from L1000 for 978 landmark genes, Gene Ontology’s expression value or gene descriptors for the perturbagen, for 4,439 genes and lastly 166-bit MACCS structural key. Giving us a total of 5,583 feature set for 20,413 small molecules.

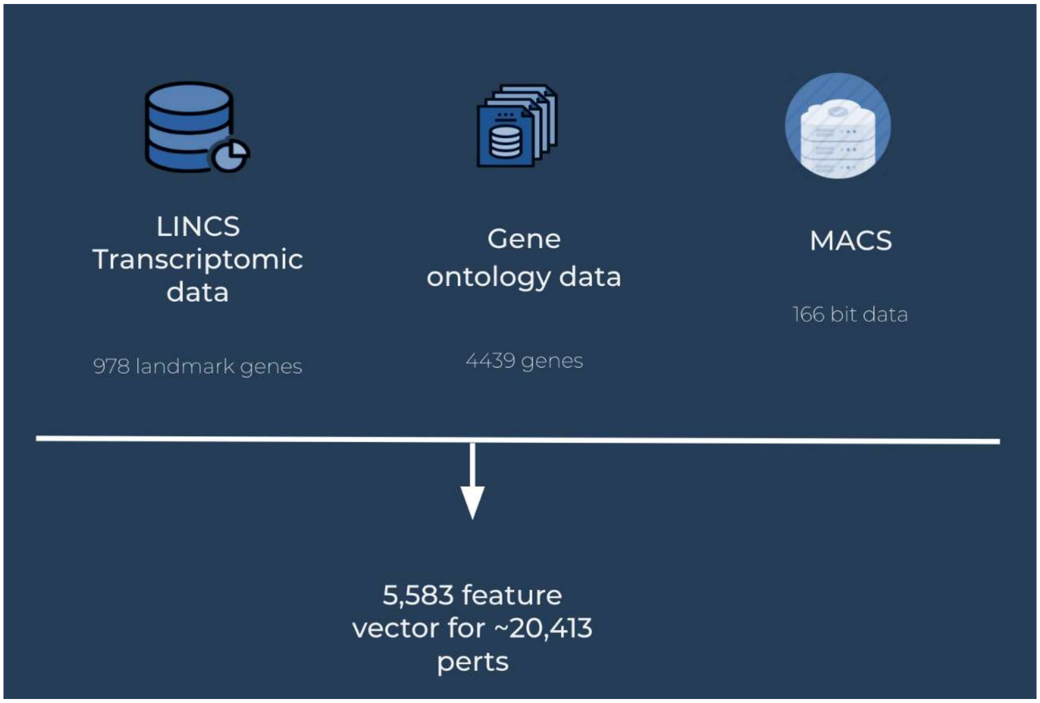
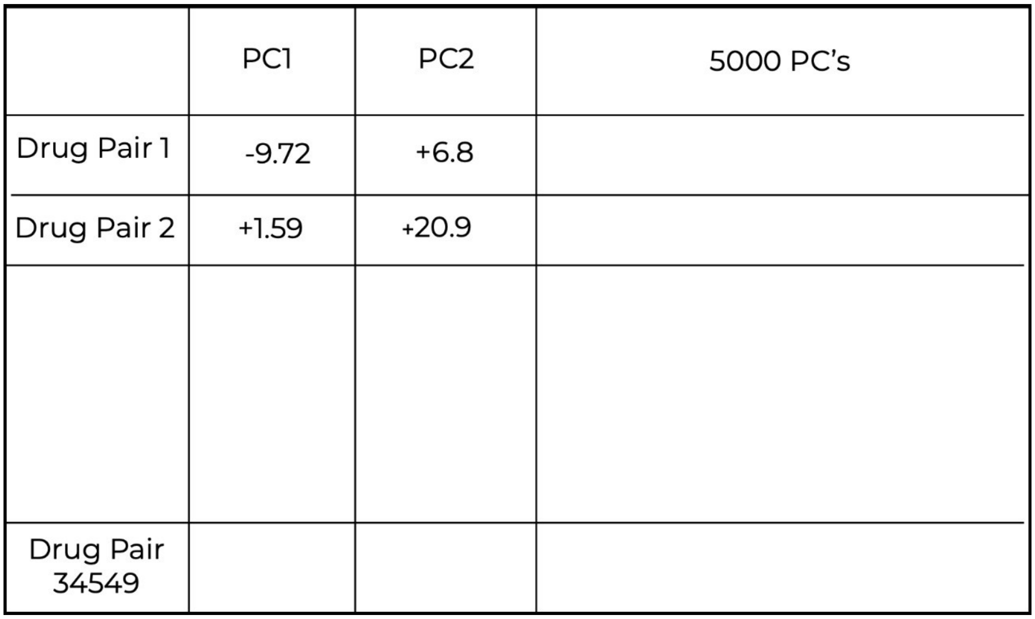

LINCS Transcriptomic data
978 landmark genes
Gene ontology data
4439 genes
MACS
166 bit data
5,583 feature vector for ~20,413 perts

Figure 3: Combined feature set for ADR prediction



	PC1	PC2	5000 PC’s
Drug Pair 1	-9.72	+6.8	
Drug Pair 2	+1.59	+20.9	
Drug Pair 34549			

Figure 4: Representation of the combine feature set2

4.5 Classification Systems

4.5.1 ATC: Anatomical Therapeutic Chemical Classification [18]

The Anatomical Therapeutic Chemical (ATC) classification system as a measuring unit are recommended by the WHO for drug utilization studies. In the Anatomical Therapeutic Chemical (ATC) classification system, the active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

Drugs are classified in groups at five different levels. The drugs are divided into fourteen main groups (1st level), with pharmacological/therapeutic subgroups (2nd level). The 3rd and 4th levels are chemical/pharmacological/therapeutic subgroups and the 5th level is the chemical substance. The 2nd, 3rd and 4th levels are often used to identify pharmacological subgroups when that is considered more appropriate than therapeutic or chemical subgroups.

The complete classification of metformin[3] illustrates the structure of the code:

A
Alimentary tract and metabolism
(1st level, anatomical main group)

A10
Drugs used in diabetes
(2nd level, therapeutic subgroup)

A10B

[2] PC = Principal components
[3] https://www.whocc.no/filearchive/publications/1_2013guidelines.pdf

Blood glucose lowering drugs, excl. insulins
(3rd level, pharmacological subgroup)

A10BA
Biguanides
(4th level, chemical subgroup)

A10BA02
metformin
(5th level, chemical substance)

Thus, in the ATC system all plain metformin preparations are given the code A10BA02.

The main groups of the ATC classification system are listed below.

A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito urinary system and sex hormones
H	Systemic hormonal preparations, excl. sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

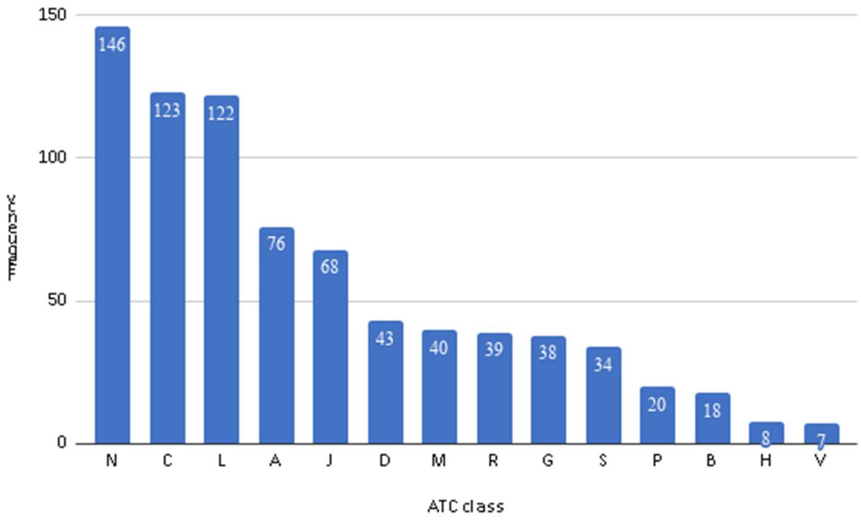


Figure 5: Distribution of ATC classes in LINCS perturbagens

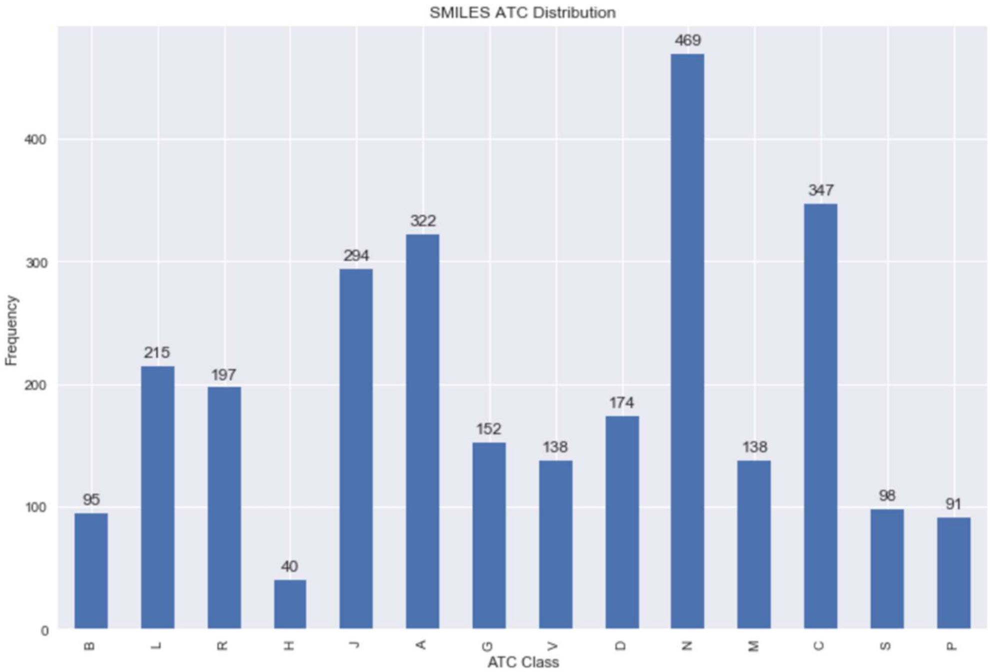


Figure 6: Distribution of ATC classes in SMILES perturbagens

4.5.2 SIDER: Side Effects Resource [19]

Unwanted side effects of drugs are a burden on patients and a severe impediment in the development of new drugs. At the same time, adverse drug reactions (ADRs) recorded during clinical trials are an important source of human phenotypic data. It is therefore essential to combine data on drugs, targets and side effects into a more complete picture of the therapeutic mechanism of actions of drugs and the ways in which they cause adverse reactions.

4.5.3 Adverse effects combined dataset

Drug Adverse Reaction (ADR) - Drug associations were extracted from two sources:

Side Effect Resource (SIDER): Stores information from the public documents about the adverse effect of the marketed drugs. At present SIDER contain data for 1430 drugs, 5868 side effects and 1396 drug-ADR association.

PharmGKB: From PharmGKB we used the database of drug-drug interaction side effects (Twosides). This database contains 868,221 putative drug-interaction effects from 59,220 pairs of drugs and 1,301 adverse events

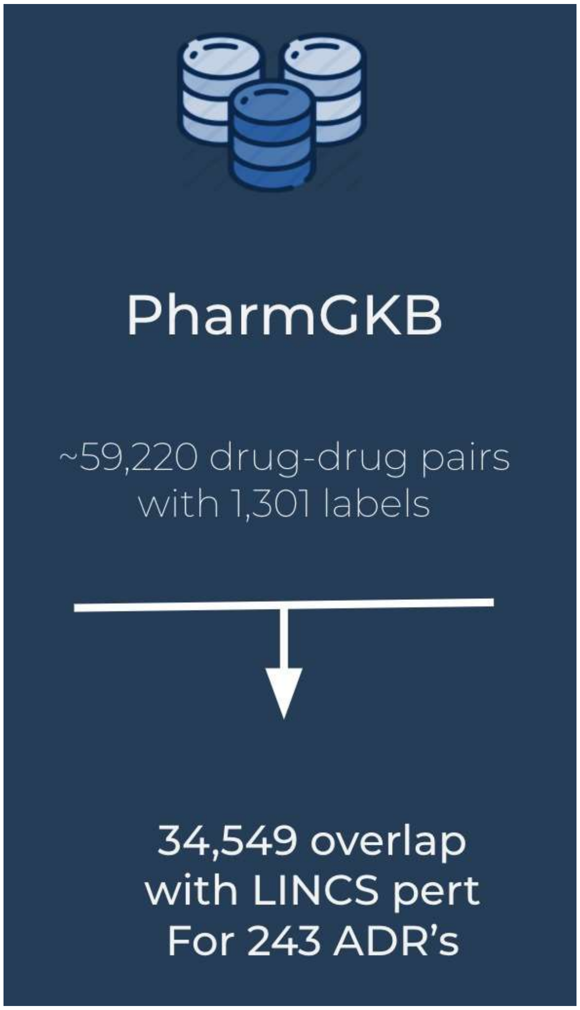

PharmGKB
~59,220 drug-drug pairs with 1,301 labels
34,549 overlap with LINCS pert For 243 ADR's

Figure 7: Combined label set for ADR prediction

Combining the above datasets with the feature set (mentioned in Section 4.4 Combined Feature Set for ADR Detection):

From the 59,220 pairs of drugs present in PharmGKB, only 34,549 drug pairs had transcriptomic information available in LINCS (60% overlap). From the 59,220 pairs of drugs present in PharmGKB, only 34,549 drug pairs had transcriptomic information available in LINCS (60% overlap).

The 5,583-feature set for each perturbagen in a pair was concatenated to get 11,166 combined feature set. PCA was then performed on this and the total number of features were brought down to 5000 for the pair.

The 1,301 adverse events were also reduced down to 243 most commonly occurring events.

Result: Dataset with 34,549 drug pairs and 5000 features, along with 243 drug labels

CHAPTER 5

PROJECT REQUIREMENTS SPECIFICATION

CHAPTER 6

SYSTEM REQUIREMENTS SPECIFICATION

CHAPTER 7

SYSTEM DESIGN

CHAPTER 8

IMPLEMENTATION AND PSEUDOCODE

CHAPTER 9

CONCLUSION OF CAPSTONE PROJECT PHASE-1

CHAPTER 10

PLAN OF WORK FOR CAPSTONE PROJECT PHASE-2

REFERENCE / BIBLIOGRAPHY

[1] Z. Xu, S. Wang, F. Zhu and J. Huang, "Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery," 2017.

[2] M. J. Keiser, "Predicting New Molecular Targets for Known Drugs," Nature, 2009.

[3] M. . Campillos, M. . Kuhn, A.-C. . Gavin, L. J. Jensen, L. J. Jensen and P. . Bork, "Drug target identification using side-effect similarity," Science, vol. 321, no. 5886, pp. 263-266, 2008.

[4] G. B. Goh, N. O. Hodas, C. . Siegel and A. . Vishnu, "SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties.," arXiv: Machine Learning, vol. , no. , p. , 2017.

[5] Y.-F. Zhang, X. Wang, A. C. Kaushik, Y. Chu, X. Shan, M.-Z. Zhao, Q. Xu1 and D.-Q. Wei, "SPVec: A Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction," Frontiers in Chemistry, 10 January 2020.

[6] S. . Jaeger, S. . Fulle and S. . Turk, "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition," Journal of Chemical Information and Modeling, vol. 58, no. 1, pp. 27-35, 2018.

[7] C. R. D. Pierri and R. Voyceik, "SWeeP: Representing large biological sequences datasets in compact vectors," Scientific Reports, 9 January 2020.

[8] Y. . Donner, S. . Kazmierczak and K. . Fortney, "Drug Repurposing Using Deep Embeddings of Gene Expression Profiles," Molecular Pharmaceutics, vol. 15, no. 10, pp. 4314-4325, 2018.

[9] C.-T. Huang, "A Large-Scale Gene Expression Intensity-Based Similarity Metric for Drug Repositioning," iScience, vol. 7, 2018.

[10] J. . Du, P. . Jia, Y. . Dai, C. . Tao, Z. . Zhao and D. . Zhi, "Gene2vec: distributed representation of genes based on co-expression," BMC Genomics, vol. 20, no. 1, p. 82, 2019.

[11] A. . Aliper, S. M. Plis, A. . Artemov, A. . Ulloa, P. . Mamoshina and A. . Zhavoronkov, "Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data," Molecular Pharmaceutics, vol. 13, no. 7, pp. 2524-2530, 2016.

[12] E. . Asgari, M. R. K. Mofrad and M. R. K. Mofrad, "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics," PLOS ONE, vol. 10, no. 11, p. , 2015.

[13] F. . Napolitano, Y. . Zhao, V. M. Moreira, R. . Tagliaferri, J. . Kere, M. . D'Amato, D. . Greco and D. . Greco, "Drug repositioning: a machine-learning approach through data integration," Journal of Cheminformatics, vol. 5, no. 1, pp. 30-30, 2013.

[14] L. . Yang and P. K. Agarwal, "Systematic Drug Repositioning Based on Clinical Side-Effects," PLOS ONE, vol. 6, no. 12, p. , 2011.

[15] Y. . Zheng, H. . Peng, X. . Zhang, Z. . Zhao, J. . Yin and J. . Li, "Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases," BMC Bioinformatics, vol. 19, no. 19, p. 517, 2018.

[16] S. . Dey, H. . Luo, A. . Fokoue, J. . Hu and P. . Zhang, "Predicting adverse drug reactions through interpretable deep learning framework," BMC Bioinformatics, vol. 19, no. 21, p. 476, 2018.

[17] C.-S. Wang, "Detecting Potential Adverse Drug Reactions Using a Deep Neural Network Model," Journal of Medical Internet Research, 6 February 2019.

[18] "Anatomical Therapeutic Chemical Classification," [Online]. Available: https://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/.

[19] [Online]. Available: http://sideeffects.embl.de.

[20] M. Abadi, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[21] G. Landrum, "RDKit: Open-source cheminformatics," [Online]. Available: http://www.rdkit.org.

[22] "Drugbank," [Online]. Available: https://www.drugbank.ca.

[23] "The LINCS Consortium," [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138.

APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS

APPENDIX B USER MANUAL (OPTIONAL)