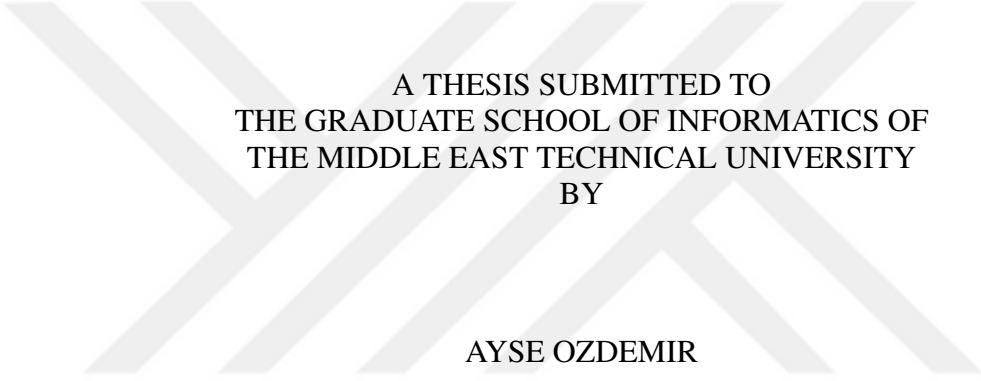


GENOTYPE CATALOG FOR THE ANALYSIS OF
DRUG-DRUG INTERACTIONS



A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AYSE OZDEMIR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

SEPTEMBER 2021

GENOTYPE CATALOG FOR THE ANALYSIS OF DRUG-DRUG INTERACTIONS

Submitted by **AYSE OZDEMIR** in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics Department, Middle East Technical University** by,

Date: 07.09.2021



I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: AYSE OZDEMIR

Signature :

ABSTRACT

GENOTYPE CATALOG FOR THE ANALYSIS OF DRUG-DRUG INTERACTIONS

OZDEMIR, AYSE

M.S., Department of Health Informatics

Supervisor: Assist. Prof. Dr. Aybar Can Acar

September 2021, 61 pages

Polypharmacy is an essential practice in today's therapeutics, especially in the care of older population. Most polypharmacy-induced drug-drug interactions (DDIs) are often discovered after drugs are put on the market. Health problems and economic burden due to unpredicted DDIs put the health system in a difficult situation. Therefore, increasing the predictability of DDIs has become one of the most critical concerns towards improving treatment success. Being dependent on several underlying parameters makes DDIs challenging to foresee. One of the most significant determinants of these underlying parameters is genetic variability. Therefore, a more profound knowledge of the DDI-genetic relationship, called drug-drug-gene interactions (DDGI), will improve treatment success. This study aims to design a relational database named DDGICat, which was designed to contribute to the ongoing DDGI research by serving as a guideline to prescribers, researchers, and the pharmaceutical industry. The content of the DDGICat was derived from other knowledge bases, including Drug-Bank, PharmGKB, Ensembl, KEGG Drug, and ONC High. DDGICat contains drugs, drug target proteins, drug-associated SNPs, DDIs, drug-gene interactions (DGIs), and DDGIs. Additionally, a developed web portal named DDGICat Browser provides the

results of mentioned content in both tabular and graphical formats. Furthermore, the end products of this study are tested in a case study on a chosen disease.

Keywords: drug-drug interaction (DDI), drug-gene interaction (DGI), drug-drug-gene interaction (DDGI), database, Single Nucleotide Polymorphism (SNP)



ÖZ

İLAÇ-İLAÇ ETKİLEŞİMLERİ ANALİZİ İÇİN GENOTİP KATALOĞU

OZDEMIR, AYSE

Yüksek Lisans, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi. Aybar Can Acar

Eylül 2021, 61 sayfa

Polifarmasi, günümüz tedavi biliminde, özellikle yaşlı nüfusun bakımında önemli bir uygulamadır. Polifarmasinin neden olduğu çoğu ilaç-ilaç etkileşimleri (DDI'ler), genellikle ilaçlar piyasaya sürüldükten sonra fark edilmektedir. Öngörülemeyen DDI lardan kaynaklanan sağlık problemleri ve ekonomik yük, sağlık sistemini zor durumda bırakmaktadır. Bu nedenle, DDI'lerin öngörülebilirliğini artırmak, tedavi başarısını iyileştirmeye yönelik en önemli amaçlardan biri haline gelmiştir. Birçok parametreye bağımlı olmaları, DDI'ların tahmin edilebilirliklerini zorlaştırmaktadır. Genetik çeşitlilik altta yatan parametrelerin en önemlilerinden biridir. Bu nedenle, ilaç-ilaç-gen etkileşimleri (DDGI) olarak adlandırılan, DDI-genetik arası ilişki hakkındaki daha derin bir bilgi, tedavi başarısını artıracaktır. Bu çalışma, reçete yazanlara, araştırmacılara ve ilaç endüstrisine kılavuzluk ederek, devam eden DDGI çalışmalarına katkıda bulunmak üzere tasarlanmış DDGICat adlı ilişkisel bir veritabanı tasarlamayı amaçlamaktadır. DDGICat'ın içeriği DrugBank, PharmGKB, Ensembl, KEGG Drug, ve ONC High dahil olmak üzere diğer veri tabanlarından türetilmiştir. DDGICat ilaçları, ilaç hedef proteinlerini, ilaçla ilişkili SNP'leri, DDI'ları, ilaç-gen etkileşimlerini (DGI'lar) ve DDGI'ları içermektedir. İlave olarak, bahsi geçen içerikleri hem tablo hem de grafik formatında sunmak amacıyla DDGICat Browser isimli bir web portalı

geliştirilmiştir. Ayrıca, bu çalışmanın nihani ürünleri, seçilen bir hastalık üzerine bir vaka çalışmasında test edilmiştir.

Anahtar Kelimeler: İlaç-ilaç etkileşimi (DDI), ilaç-gen etkileşimi (DGI), ilaç-ilaç-gen etkileşimi (DDGI), veritabanı, Tek Nükleotid Polimorfizmi (SNP)





To my mother, Makbule Koç
Annem, Makbule Koç'a

ACKNOWLEDGMENTS

First and foremost, I would like to express my deep and sincere gratitude to my research supervisor, Assist. Prof. Dr. Aybar Can Acar. His constant support, insightful comments, and friendly attitude allowed this study to complete. It was a great privilege and honor to study under his guidance.

Besides, I would like to thank Assoc. Prof. Dr. Yeşim Aydın Son and Assoc. Prof. Dr. Tunca Doğan for reviewing my work.

Special thanks to my mother, Makbule Koç, who has supported me with unconditional love throughout my life. This dissertation would not have been possible without her.

I would like to thank my siblings, Zeynep Özsubaşı, Hasan Özdemir, and Özlem Özdemir, who have encouraged and supported me through my graduate education. Additionally, I would like to thank my nephews Miray, Nisa, and Gökçe for their patience as they wait to finish my studies.

I would like to thank my dear best friend, Hüseyin Demir. Hüseyin has supported and encouraged me with all his heart since the first day we met.

I would like to thank my dear friend Zehra Demirtaş, who has supported me with her invaluable guidance through my graduate education. I am grateful to her for never stop believing that I will succeed.

I would like to thank Emine Çimen for supporting me with her valuable comments from the beginning of my graduate journey.

I would like to thank Alejandro Gonzalo for his constant motivational support and for being very generous with his time in listening to my endless thesis stories.

TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZ.....	vi
DEDICATION.....	viii
ACKNOWLEDGMENTS	ix
TABLE OF CONTENTS	x
LIST OF TABLES	xiii
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS	xvi
CHAPTERS	
1 INTRODUCTION	1
1.1 Motivation	1
1.2 Goal	2
1.3 Thesis Outline.....	2
2 RELATED WORK	5
2.1 Drug: From Past to Present	5

2.2	Pharmacology	6
2.2.1	Pharmacokinetics	7
2.2.2	Pharmacodynamics	8
2.3	Drug-Drug Interactions (DDIs)	8
2.4	Drug-Gene Interactions (DGIs)	9
2.5	Drug-Drug-Gene Interactions (DDGIs)	10
2.6	Drug Knowledge Bases Used	10
2.6.1	DrugBank	11
2.6.2	PharmGKB	11
2.6.3	Ensembl	12
2.6.4	DDI Severity Sources	13
2.7	Related Studies	13
2.7.1	DDI & DGI Studies	14
2.7.2	DDGI Studies	14
3	MATERIALS AND METHODS	17
3.1	Data Sources	17
3.1.1	DrugBank	17
3.1.2	PharmGKB	19
3.1.3	Ensembl	19
3.1.4	DDI Severity	20

3.2	Data Integration	21
3.3	Data Preprocessing	22
3.4	Database Creation	23
3.5	Querying DDGICat	30
4	RESULTS	37
4.1	Case Study	37
4.2	Database Statistics	41
5	DISCUSSION & CONCLUSION	51
5.1	Importance of Drug-Drug Interactions	51
5.2	Importance of Drug-Drug-Gene Interactions	51
5.3	Importance of DDGICat	52
5.4	Significance of the Data in DDGICat	52
5.5	Conclusion	54
5.6	Future Studies	54
	REFERENCES	55
	APPENDICES	
A	PROJECT CODES	61

LIST OF TABLES

Table 1	Previous DDI, DGI, and DDGI studies	16
Table 2	Drug statistics a) Drug counts per drug type b) Drug counts per drug group c) Drug counts per combination of drug type and group.	18
Table 3	Drug-protein statistics a) Grouped by drug-protein b) Grouped by drug-protein and drug type.....	18
Table 4	PharmGKB data files a) Clinical annotations b) Relationships	19
Table 5	biomaRt parameters used to query data from Ensembl.	20
Table 6	Severity information of interacted drug pairs retrieved from KEGG Drug and ONC High (CI: Contraindication, P: Precaution)	21
Table 7	Statistical summary of DDGICat entities	21
Table 8	Interacting drug pair classification based on interaction description information (The descriptions are taken verbatim from DrugBank)	50

LIST OF FIGURES

Figure 1	Data flows used in data integration	22
Figure 2	ER Diagram of DDGICat	28
Figure 3	The Relational Data Model of DDGICat (Screenshot from PgAdmin 3)	29
Figure 4	Screenshot of Downloads Page on DDGICat Browser	30
Figure 5	Screenshot of Drug Page on DDGICat Browser	31
Figure 6	Screenshot of Gene and SNP Pages on DDGICat Browser	32
Figure 7	Screenshot of DDI and DGI Pages on DDGICat Browser	33
Figure 8	Screenshot of DDGI Page on DDGICat Browser. The DDGI Page shows the relation between a disease and a drug pair in SNP, protein, gene, and chromosome detail	34
Figure 9	Data retrieval logic behind Drug, Gene, SNP, DDI, DGI, and DDGI Pages.	35
Figure 10	Drugs and drug-associated proteins interacting with Clopidogrel. (Screenshot of DDI and DGI Pages)	39
Figure 11	Severity of Clopidogrel and Omeprazole interaction affected by genetic materials. (Screenshots of DDGI Page)	40
Figure 12	Drug types classified per drug status	41
Figure 13	Drug-related gene distribution on chromosomes	42
Figure 14	Distribution of drug-related SNPs on chromosomes	43
Figure 15	Top 10 genes interacting with most drugs	44
Figure 16	Distribution of drug-interacting protein types	44
Figure 17	Top 10 drug-interacting proteins	45

Figure 18	Drug-protein distribution per drug	46
Figure 19	DDI distribution per drug	47
Figure 20	DDI distribution per drug-protein	48
Figure 21	DDI Percentages per drug-protein	49
Figure 22	DDI distribution per ATC level	49



LIST OF ABBREVIATIONS

ADR	Adverse Drug Reaction
CDST	Clinical Decision Support Tool
CYP-450	Cytochrome P450
DDI	Drug-Drug Interaction
DGI	Drug-Gene Interaction
DDGI	Drug-Drug-Gene Interaction
EM	Extensive Metabolizer
EMA	European Medicines Agency
GIT	Gastrointestinal Tract
IM	Intermediate Metabolizer
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
PharmGKB	The Pharmacogenomics Knowledge Base
PD	Pharmacodynamics
PK	Pharmacokinetics
PM	Poor Metabolizer
RDBMS	Relational Database Management System
SNP	Single Nucleotide Polymorphism
UM	Ultra-rapid Metabolizer
Vd	Volume of Distribution

CHAPTER 1

INTRODUCTION

1.1 Motivation

Emergency department visits and rehospitalizations due to Adverse Drug Reactions (ADRs) are a threatening problem for public health [1]. Drug-drug interaction (DDI) is a particular case of ADRs and occurs when a drug alters the effect of a concomitant drug. As stated in a study prepared by evaluating the results of 23 studies worldwide, DDIs cause 0.054% of emergency room visits, 0.57% of hospital admissions, and 0.12% of rehospitalizations [2].

Drug-gene interaction (DGI) explains drug phenotype differences among individuals caused by genes and genetic variances. DGI has started to gain more attention after realizing that not every drug has the same effect on everybody. As reported by research conducted among 10,000 patients, more than 90% of each patient had at least one genetic variation, leading to drug response differences [3].

Drug-drug-gene interaction (DDGI) is a subset of DDI, and it is the cumulative effect of DDI and DGI. According to a study conducted among 30 patients, approximately one-third of patients had DDGIs [4].

Although genetic variability on DDI occurrence is apparent, it still does not receive the deserved attention. However, with the contribution of the fact that genetic testing has become more accessible at lower costs, the effect of genetic variability in current DDI studies has started to be considered more prevalently [5].

Most DDGIs are being detected after drugs are put on the market during post-marketing surveillance. This unpredictability is a threatening risk for public health; therefore, early detection of DDGIs could reduce the economic and health burden on both patients and the pharmaceutical industry.

1.2 Goal

Due to recent technological developments, the drug development process has become both less costly and more error-free. One of the essential instruments contributing to this success is the convenience of accessing the pharmacological data.

Similarly, the increasing amount of biological, pharmacological, and pharmacogenomics (PGx) data is significant for ongoing DDI, DGI, and DDGI research. Unfortunately, gathering this data is challenging since it is primarily found in separate sources in a heterogeneous state [6]. Another issue is that biological data found in different sources can contain contradictions or have minimal overlap with each other [6]. The situation is also the same with PGx data. Currently, most DDI databases do not involve genotype focussed data. Due to the mentioned challenges, ongoing DDI, DGI, and DDGI studies are being hampered.

This study aims to contribute to the mentioned gap in the literature by providing a combined catalog database named DDGICat, which was developed by extracting information from several knowledge bases. DDGICat consists of the derived data on drugs, genes, drug-associated genes, DDI, DGI, and DDGI. Unlike previous research, DDGICat provides information on the DDI-genetic relationship. Moreover, a web interface named DDGICat Browser allows viewing DDGICat content and statistical visuals through a web portal.

Furthermore, for future studies, the content of the DDGICat may serve as a guideline in predicting DDIs, DGIs, and DDGIs that have not yet been anticipated.

1.3 Thesis Outline

This thesis comprises five parts, and the outline is as follows.

- Chapter 2 gives background information on DDIs, DGIs, DDGIs, and the chronological drug and pharmacology overview. Towards the end of this chapter, a critique of related studies and used data sources is demonstrated.
- Chapter 3 describes technical details about the applied data integration and data pre-processing steps to construct DDGICat. Similarly, technical details of the developed web portal, which enables user interaction with DDGICat, are demonstrated.
- Chapter 4 provides sample statistics retrieved from DDGICat. Through the end of this chapter, a sample use case scenario demonstrating an example for the real-life usage of the developed catalog (DDGICat) is presented.

- Chapter 5 mentions the importance of drug-drug interaction, drug-drug-gene interaction, and DDGICat. Moreover, the significance of the data in DDGICat is discussed. Additionally, this chapter reviews accomplished tasks and gives recommendations for future studies.





CHAPTER 2

RELATED WORK

This chapter puts forth the following topics in order: Section 2.1 reveals a historical overview of drugs, which is the primary subject of this study. Section 2.2 introduces a brief pharmacology overview that will assist the reader through this study. Section 2.3 explains drug-drug interactions (DDIs). Section 2.4 describes drug-gene interactions (DGIs). Section 2.5 presents drug-drug-gene interactions (DDGIs). Section 2.6 introduces the knowledge bases used in this research. Finally, Section 2.7 summarizes previous studies in the literature.

2.1 Drug: From Past to Present

A drug that is the principal product in pharmacology takes its origin from the French word fragrance (Droque) with the meaning of dry herb. According to the definition of the WHO (World Health Organization), “Drug is any substance or product that is used or is intended to be used to modify or explore physiological systems or pathological states for the benefit of the recipient” [7].

One of the most challenging problems of humankind has been to find novel ways to advance health and longevity. Ötzi, who was called the iceman and lived around 3200 BC, also had the same ambition. The fungus found in his belongings was thought to cure parasitic worms, proving his ambition for wellness. Based on this proof, it would not be wrong to say that the history of drugs used for medicinal purposes is as old as human history.

This ambition for wellness has been a cultural heritage transferred among civilizations, including Sumerians, Egyptians, Indians, Chinese, Greeks, Roman Empire civilization, and Arabs [8]. The famous saying of Hippocrates, “Primum non nocere!” meaning “first, do not harm” is one of the main rules taught in medical schools today [9].

2.2 Pharmacology

The word *pharmakon* and *-logia* both come from Greek (Pharmacology, n.d.). Pharmacology is a combination of these two words [8], meaning the science that studies the interaction of living systems with chemical molecules [10].

Rudolf Buchheim, known as the founding father of pharmacology, established the first institute of pharmacology, in 1847, at the University of Dorpat, in Germany [11]. In its initial periods, pharmacology was applied by trial and error and progressed by finding the correct answers to which drug, which dose, how often, and how long. The process that involves finding the most accurate answers to these questions is called the Drug Development Process.

The drug development process involves identifying the abnormality that causes the disease and, subsequently, finding the most efficient therapeutic molecule to fix or mitigate this abnormality. These steps are mainly conducted by *in vitro* and *in vivo* studies. *In vitro* studies are performed in a test tube in a laboratory environment without using living organisms. *In vivo* studies contain experiments with living organisms [12]. The inclusion of *in silico* studies based on software technologies into the drug development process has significantly decreased time and costs [13].

After finding the most therapeutic molecules to fix or mitigate the abnormality, a new step named the Preclinical Development process starts. This step includes laboratory studies on animals. Clinical Trials are the subsequent step and involve five consecutive phases [14]. Phase 0 consists of experiments with a low dose of the drug on healthy volunteers. Phase 1 contains tests conducted on healthy individuals. Phase 2 includes studies with patients who have the disease. Phase 3 holds trials with more patients who have the disease. Finally, phase 4 comprises observations after a drug is put on the market.

The shortening of the time required for the drug development process has increased the number of drugs produced. Therefore the growing number of drugs has created the need to classify them, and various methods have been developed. One of these methods is ATC (Anatomical Therapeutic Chemical). WHO controls ATC, and it uses pharmacological, therapeutic, and chemical properties to classify the drugs into five groups hierarchically (one leading group and four hierarchical subgroups). Similarly, another classification is viable to drug categories. Drugs are mainly divided into two main categories, including biotech and small molecule drugs. While biotech drugs are extracted from living organisms, small molecule drugs are obtained from various chemicals. Classifying drugs according to their names is an alternative way since drugs are named differently for different purposes. The first one is the Chemical Name, which has the function of specifying the drug chemically. The second one is the Generic Name chosen by the government, and the third one is the Brand Name

given by the manufacturer, which is the only agent holding the right to produce and sell the drug within a certain period. Pharmacology has two subgroups, including Pharmacokinetics and Pharmacodynamics, and the following sections provide their details.

2.2.1 Pharmacokinetics

The word kinesis comes from Greek and means movement. Pharmacokinetics (PK) is a subfield of pharmacology and analyzes “what the body does to the drug” [15]. PK has four drug-related subgroups referred to as ADME (Absorption, Distribution, Metabolism, Excretion).

Absorption (A) explains how drugs move from the site of administration into the bloodstream [16]. Depending on the route of administration, drugs have different absorption mechanisms. For instance, while intravascularly administered drugs do not need absorption as they are already in the bloodstream, orally administered drugs are absorbed through the GI (gastrointestinal) system and sent to the liver and bloodstream, respectively [17].

Distribution (D) defines the movement of drugs through the body to reach their site of action (Seifert, 2019). During drug distribution in the body, many drugs bind plasma proteins, and this chemical reaction is called plasma protein binding (Seifert, 2019). However, plasma protein binding may lead to undesired results since it prevents access to the primary target that the drug should activate [17]. The volume of distribution (Vd) is another related term that measures how quickly a given dose of the drug reaches the required therapeutic plasma concentration [14].

Metabolism (M) refers to transforming a drug into a more water-soluble form by the liver. Drug metabolism consists of two stages, including phase I and phase II. While Phase I reactions convert the drug into a more water-soluble form, Phase II reactions are mainly responsible for drug elimination [17]. Bioavailability defines the percentage of administered drugs included in systemic circulation in the unchanged form. For example, while intravenously administered drugs offer nearly 100% bioavailability, orally administered ones provide less bioavailability since a certain amount is eliminated due to metabolization in the liver [16]. This process is called the first-pass metabolism. Therefore, bioavailability and first-pass metabolism are the terms used to describe drug metabolism.

Each organism has its metabolic rate depending on the polymorphism of the metabolic enzymes. Based on the polymorphism of metabolic enzymes, an organism is classified into poor (PM), intermediate (IM), rapid (RM), and ultra-rapid (URM) metabolizers. Since the metabolizing enzymes are less active, PMs or IMs metabolize drugs

rather slowly, needing lower doses to prevent undesired toxicity. On the other hand, RMs or URMs break down drugs quickly, and they need higher doses to get the desired effects.

Excretion (E) is defined as eliminating drugs from the body mainly by the liver and kidneys. Half-life and clearance are the terms used for the measurements of this step. Half-life measures the required time for the drug to drop to its half amount in the body. Clearance defines the ability of the body to eliminate the drug.

2.2.2 Pharmacodynamics

The word dynamics comes from Greek and means power. Pharmacodynamics (PD) is a subfield of pharmacology and analyzes “what the drug does to the body”. In other words, PD focuses on how drugs act on predefined biochemical reactions at their targets [18].

Drugs generally interact with four different types of protein, including target, enzyme, carrier, and transporter. Targets, also called receptors, are macromolecules that recognize the drug and initiate the response. Agonist and antagonist terms are used to describe drug and receptor interaction. In agonist reactions, drugs cause the expected physiological response by attaching the predefined receptor, whereas, in antagonist reactions, drugs prevent the expected physiological response by blocking the receptor [19]. Enzymes are mainly responsible for drug metabolism. Carriers, also called ion channels, regulate drug flow across cell membranes. Finally, transporters transport substrates while entering or leaving cells [10].

Toxicity (T) is another critical term for drug pharmacology. Toxicity refers to the situation when a drug damages a biological target. Pharmacology and toxicity intersect very commonly since the primary goal of pharmacology is to increase therapeutic effects while decreasing toxicity. Therefore ADMET (Absorption, Distribution, Metabolization, Excretion, Toxicity) is an interchangeable term of ADME.

2.3 Drug-Drug Interactions (DDIs)

Polypharmacy refers to using multiple drugs concomitantly, and it is a part of today’s therapeutics. Although some polypharmacy-induced effects are desirable for the sake of treatment, in most cases, they cause harm. The term Adverse Drug Reactions (ADRs) is a technical term defining these detrimental effects.

The probability of ADR is positively correlated with the number of concomitant drugs. According to a cohort study, a patient taking 5-9 medications has a 50%

risk of having ADR, whereas a patient taking 20 or more medications has a 100% risk of having ADR [20]. Since the number of simultaneous diseases increases in advanced ages, the elderly are inevitably more vulnerable to polypharmacy-induced ADRs [20], [21], [22]. One of the most common causes of ADRs is drug-drug interactions (DDIs) [23].

DDI refers to the interaction between two or more co-administered drugs when one drug changes the effect of another drug in both a desirable or undesirable way. A well-known example of a desirable DDI is the Probenecid and Penicillin pair. Probenecid is given together with Penicillin to decrease the excretion of Penicillin for a longer duration of action in the body [24]. In contrast to desirable DDIs, undesirable DDIs led to many health and economic problems. For example, according to an analysis based on 23 clinical studies, DDIs cause 0.054% of emergency room visits, 0.57% hospital admissions, and 0.12% rehospitalizations [2].

Although anticipating DDIs beforehand is essential in most cases, due to the hardship of detecting all drug combinations experimentally, DDIs are often realized during post-marketing surveillance after drugs are sent to the market [25]. Emerging data-oriented *in silico* approaches seem to have considerable potential in the solution of the problem. With the help of *in silico* techniques, predicting the hazardous effects of DDIs at the time of prescription seems to prevent unexpected results. A study conducted among the elderly in polypharmacy revealed that CDST (Clinical Decision Support Tool) used in the study reduced re-hospitalization and emergency department visits [26].

2.4 Drug-Gene Interactions (DGIs)

In 2006, a Codeine prescribed breastfeeding mother lost her baby due to morphine overdose, although other breastfeeding mothers having the same prescription did not have any problem with their babies [27]. The underlying reason for the baby's death was the polymorphism on the CYP2D6 enzyme of the mother, which led her to metabolize Codeine at a relatively fast rate (URM). It is soon discovered that an excessive amount of morphine, which is the product of Codeine metabolism, passed into the baby through the mother's breast milk.

In 1957, Arno Motulsky first put forward that genetic differences may lead to drug-response variability [28]. 2 years later, Friedrich Vogel coined the term Pharmacogenetics [29]. Afterward, Marshall coined the term Pharmacogenomics in 1997 [30]. Pharmacogenetics focuses on different drug responses among individuals having different genetic makeup. Similarly, Pharmacogenomics explores all genes in the genome. Therefore, PGx is a technical term referring to Pharmacogenetics and Pharmacogenomics.

With recent advances in PGx studies, drug-gene interactions (DGIs) gained popularity in measuring the effect of genetic variability on drug response [31]. As stated in a study, genetics explains variability in drug response by 20–95 % [29]. Similarly, inter-individual variability of CYP450 genes explains 25% of the variability in drug response [32].

Considering the substantial effect of DGIs on drug response, incorporating them to predict DDIs may increase the prediction accuracy [33]. For instance, in a study [34] the drug-gene interaction information has been used to estimate the unknown drug-drug interactions, and approximately 80% of the prediction success has been achieved.

2.5 Drug-Drug-Gene Interactions (DDGIs)

Various similarity approaches such as chemical, biological, target, functional, side effect, and metabolism of drug pairs have been used to reveal undetected DDIs [35], [36], [37], [38], [39]. In addition, several computerized techniques, including classification, clustering, text mining, and graph simulations, led to unknown DDIs being predicted [40], [34], [39], [41], [38]. Moreover, following the recent advances in PGx studies, DDIs also have started to be reviewed from the perspective of genetics [42]. For example, a genotype-guided DDI study among the elderly in polypharmacy revealed that not only emergency department visits and re-hospitalization but treatment costs also decreased [43]. Similarly, revealing CYP-450 mediated DDIs has enabled predicting and reducing DDIs in the hospitalized elderly [44].

Drug-drug-gene interaction (DDGI) is the cumulative effect of DDI and DGI since the aggregate effect of concomitant drug usage and genetic variability alters predicted drug response [45].

Although there is no simple explanation or generated model of DDGIs, their profiles, including the order of administration, route of administration, dose, and genotype of the metabolizing enzymes, are feasible [46]. Furthermore, a literature review searching the relation of DDIs with major metabolizing enzymes, including CYP2C9, CYP2C19, and CYP2D6, revealed that polymorphisms on these metabolizing enzymes are an essential determinant of DDGIs [47].

2.6 Drug Knowledge Bases Used

This section details the drug knowledge bases used in this study: DrugBank, PharmGKB, Ensembl, KEGG Drug, and ONC High Priority. Drug information, which is

the primary material of this study, is obtained from DrugBank since it is a comprehensive drug encyclopedia. In addition, drug-associated gene and SNP records are extracted from PharmGKB and Ensembl. Furthermore, DDI records retrieved from DrugBank are fed by the severity information gathered from KEGG Drug and ONC High Priority.

2.6.1 DrugBank

DrugBank [48], also known as a drug encyclopedia, was first published in 2006. It is a detailed, blended bioinformatics and cheminformatics drug database containing drug information mainly extracted from literature. DrugBank has a wide range of users, including researchers, students, physicians, pharmacists, medicinal chemists, and the pharmaceutical industry. DrugBank content has open access with a membership prerequisite.

DrugBank has information on drugs, drug chemical properties, ADMET details, druggable genes, drug proteins, drug pathways, drug-SNP associations, DDIs, and drug-food interactions. In this study, drug, DDI, and drug-associated SNP records are extracted from DrugBank.

Drug entity has information about 225 different organisms. It has 14315 drug entries, which are composed of 2481 biotech drugs and 11834 small molecule drugs. Based on the FDA statuses, drugs are categorized into six groups: approved, investigational, experimental, nutraceutical, illicit, and vet-approved. The DDI entity has around one million drug-drug interactions. These interaction records contain clinically verified entries as well as prediction-based ones. DrugBank gathered clinically proven records mainly from drug labels and literature. In addition, it contains prediction-based entries generated by using several machine learning algorithms.

The drug-associated SNP entity has 308 records, with column names drug identifier, SNP identifier, gene, UniProt identifier, allele, defining change, and PubMed identifier, 201 of which are harmless and 107 of which are adverse interactions.

2.6.2 PharmGKB

The Pharmacogenomics Knowledge Base (PharmGKB) [49] is a publicly available online resource that was developed at Stanford University in 2000 and funded by the NIH (National Institutes of Health). PharmGKB data mainly focuses on the relations between variant, drug, and phenotype. This data has been prepared by manual curation of the literature and natural language processing techniques. PharmGKB shares its content with a free membership prerequisite in two separate parts. The first part,

also mentioned as primary data, comprises each gene, drug, variant, and phenotype definition. The second part contains curated literature records, including variant annotations, clinical annotations, drug label annotations, clinical guideline annotations, and pathways [50]. At the time of this study, PharmGKB included 712 drug annotations, 780 drug label annotations, 165 clinical guideline annotations, and 151 pathway annotations [March 2021].

The primary PharmGKB data contributing to this study are the Clinical Annotations and Relationships files. The Clinical Annotations file has information on chromosomes, gene, SNP, drug, disease, level of evidence, and PubMed identifier. Similarly, the Relationships data file holds the association information on drug, gene, SNP, disease, PK, PD, evidence, and PubMed identifier.

The Clinical Annotation file contains 4559 drug-associated SNP records. The Relationships data file contains 61604 records on drug-gene, drug-variant, drug-drug, drug-disease, gene-disease, and disease-variant associations.

Most of the PharmGKB content includes an additional rating attribute including high, moderate, low, and unsupported given by PharmGKB curators describing the evidence, which is a combinatorial value depending on the trustiness of the evidence or the number of the published paper about the case.

2.6.3 Ensembl

Towards the end of the Human Genome Project, known as the most prominent biological joint project, to meet the need of classifying, integrating, and demonstrating vast amounts of annotated genomic data, the Ensembl project started with the funding of EMBL European Bioinformatics Institute and the Wellcome Trust Sanger Institute, in 1999. Ensembl consolidates massive biological data retrieved from several sources and shares this content from a central point with a free membership prerequisite.

Ensembl content includes genes, variants, phenotypes of certain species (human, mouse, zebrafish, and rat). For the mentioned species, Ensembl imported information from the HAVANA project. Every gene in Ensembl has a unique identifier starting with the ENSG prefix. In the same way, transcripts are named with the ENST prefix. Variation data for the human genome in Ensembl consist of the imported records from dbSNP.

To solve the problem of timely access to the latest genomic data, Ensembl provides the data via several tools. Ensembl Browser is one of these tools and allows searching biological information of over 250 species, with gene name, gene symbol, gene identifier, variation, disease, and phenotype search keys.

Another tool is BioMart, a browser-based genome query tool enabling users to export cross-database gene information for selected species in different formats. Several filtering options provide users to define customized search criteria, including region, gene, phenotype, and variant. Programmatically access to the Ensembl data is another alternative. For instance, an open-source R package called biomaRt [51] enables users to retrieve the genomic content. It provides retrieval of data via predefined R objects without knowing the complete structure of the database or constructing complex SQL queries. In this research, 11,889 gene entries and the 2,093 SNP records are extracted from Ensembl via biomaRt.

2.6.4 DDI Severity Sources

DDI data extracted from DrugBank does not have severity information. Therefore, the severity information of interacted drug pairs was extracted from other sources, including KEGG Drug and ONC High Priority.

KEGG (Kyoto Encyclopedia of Genes and Genomes) [52] is an integrated knowledge base, first started in Kanehisa Laboratories in 1995. KEGG has a broad information spectrum, including health, genomic and chemical on 18 different domains, one of which is the KEGG Drug.

KEGG Drug is a comprehensive drug resource containing general drug information, PK, PD, and variant information of approved drugs in Japan, the USA, and Europe. Drug Interaction Database is a subset of KEGG Drug, and it has derived contraindication (CI) and precaution (P) data for prescription drugs available in Japan. Drug Interaction Database has 217,854 rows containing severity and description of interacted drug pairs.

The ONC High Priority [53] project was started by ONC (Office of the National Coordinator for Health Information Technology) to determine drug-drug interactions having high severity. ONC High contains information on 602 critically interacted drug pairs.

2.7 Related Studies

Drug-drug-gene interaction (DDGI) is the cumulative effect of DDI and DGI. Therefore, this section reveals previous studies of DDI, DGI, and DDGI, respectively, to evaluate the existing work holistically. Later in this section, a summary of the previous work is summarized in Table 1.

2.7.1 DDI & DGI Studies

ADReCS-Target [54] is the **Adverse Drug Reaction Classification System-Target** which contains protein, gene, and genetic variation information associated with adverse drug reactions. It was funded by Bioinformatics & Design group (BIDD). Currently, it includes more than 65000 ADR associations of 662 drugs with 63298 genes, 2613 variations, 1710 proteins. The database content is derived from literature and other knowledge bases, including DrugBank, Ensembl, and GWAS Catalog. The database has open access to download with a free membership prerequisite.

DGIdb [55] is the **Drug-Gene Interaction database**. It contains information on drug-associated genes derived from over thirty sources such as literature, clinical trial records, and other knowledge bases, including DrugBank, PharmGKB, ChEMBL, Drug Target Commons, and Therapeutic Target Database (TTD). The database has over 1,000,000 drug-gene interaction records associated with around 10,000 drugs and more than 40,000 genes. DGIdb content is accessible through the provided web portal, downloadable data files, and provided API.

Merged-PDDI [6] is the **Merged Potential Drug-Drug Interactions**. The University of Pittsburgh funded it. In total, Merged-PDDI is a synthesis of 14 different sources, including CredibleMeds, NDF-RT, ONC High Priority, ONC Non-interruptive, DDI Corpus, KEGG DDI, TWOSIDES, DrugBank. Merged-PDDI content has open access through both the provided web portal and data downloading options.

PreMedKB [56] is the **Precision Medicine Knowledge Base**, and Fudan University funded it. It contains association information on genes, diseases, drugs, and variants. PreMedKB has around 200,000 genetic variations, 29,000 drug-gene associations, and 6,000 variant-drug pairs derived from several resources, including HGNC, NCBI, UniProtKB, ClinVar, dbSNP, DailyMed, Drugs@FDA, DrugBank, PubChem, STITCH, PharmGKB, and TTD.

VarDrugPub [57] is funded by the National Research Foundation of Korea. It provides pharmacogenomic data content, which was generated by deriving information from around 6,000 PubMed papers. It has information for 901 drugs, 1,077 genes, 3,591 mutations, and 5,712 drug-variation associations.

2.7.2 DDGI Studies

DDI-Predictor [58] is an online decision-making tool assisting pharmacists with their DDI decisions. The study is ongoing and started by analyzing prescriptions by 18 clinical pharmacists by the Genophar Working Group at the University Claude Bernard Lyon. Among the 199,733 prescriptions, 213 cases are meaningful, and they

are divided into groups of inducers, inhibitors, cirrhotic patients with percentages of 26, 68, and 6, respectively. The tool has an online interface providing search options based on five modules, including drug-drug interactions, drug-gene interactions, drug-drug-gene interactions, cirrhosis-drug interactions, cirrhosis-drug-drug interactions. The DDI module provides search options on the selected drug, substrate, and inducer pairs. Similarly, the DDGI module makes predictions on the selected substrate, interactor, and genotype pairs. The data is open access through the provided API named DDPRED and downloadable data files.

[59] designed to predict detrimental DDIs between two drugs by considering genetic interaction between genes that encode these drugs' targets. In order to build the prediction model, DDI data is retrieved from other knowledge bases, including DrugBank, TWOSIDES, and Merged-PDDI. The generated dataset contains 1,113 adversely interacted drug pairs, 11,113 non-interacted drug pairs. The final model predicts 432 novel adverse DDIs and provides supporting evidence regarding the prediction results.

Table 1: Previous DDI, DGI, and DDGI studies

Name	Year	Owner	Source	Entity
ADReCS-Target	2018	Bioinformatics-Aided Drug Discovery Group (BADD)	PubMed, DrugBank, Ensembl, GWAS Catalog	ADR associated protein and gene
DGIdb	2017	Washington University	More than 30 databases, including PubMed, clinical trial records, DrugBank, PharmGKB, TTD, etc	Drug associated gene
Merged-PDDI	2017	University of Pittsburgh	14 sources, including CredibleMeds, NDF- RT, ONC, KEGG DDI, TWOSIDES, DrugBank, etc.	DDI
PreMedKB	2019	Fudan University	More than 20 databases, including DrugBank, PharmGKB, TTD, etc.	drug, gene, disease, and variant associations
VarDrugPub	2018	Korea University	PubMed, ClinVar, PharmGKB, PubTator	drug, gene, and variant associations
DDIPredictor	2019	University Claude Bernard Lyon	EHR	drug, protein, gene, variant, and DDI associations
Qian et al., 2019	2019	Cornell University	DrugBank, TWOSIDES, Merged-PDDI	Adverse DDI, gene, and protein associations

CHAPTER 3

MATERIALS AND METHODS

This chapter explains data integration and preprocessing operations with tables, visuals, the Entity-Relationship Diagram (ERD), and Relational Data Model (RDM). Moreover, through the end of this chapter, database entities are explained in detail. Finally, the developed web interface which enables users to view the end product of this study was introduced.

3.1 Data Sources

3.1.1 DrugBank

The latest published version of DrugBank at the time of this study is 5.1.8, released on 2021-01-03. DrugBank has an API to query the database content partially and a downloadable 1.5 GB sized XML-formatted full dataset, which was downloaded and used in this project. This dataset was parsed with an open-source R package called “dbparser” [60] and imported into the local RDBMS (Relational Database Management System). The parsed data has several entities, including drug, drug-interacting protein, DDI, and drug-associated SNP.

The **drug** entity has comprehensive information on name, synonym, indication, pharmacokinetics (PK), and pharmacodynamics (PD). This entity has 14,315 drugs, 11,834 are small molecules, and 2,481 biotech drugs. Drugs are categorized into six groups based on the FDA approval status: experimental, investigational, nutraceutical, illicit, vet-approved, and withdrawn. A drug might belong to more than one group since it could have been approved to treat disease while it is in the trial to treat another disease. A summary of the drug entity is given in Table 2.

Table 2: Drug statistics a) Drug counts per drug type b) Drug counts per drug group
c) Drug counts per combination of drug type and group.

a)		b)	
Type	Number of Drugs	Group	Number of Drugs
Biotech	2,481	Approved	4,108
Small Molecule	11,834	Experimental	6,554
Total	14,315	Illicit	205
		Investigational	5,245
		Nutraceutical	131
		Vet Approved	423
		Withdrawn	265
		Total	16,931

c)	
Type (Approved)	Number of Drugs
Small Molecule	2,675
Biotech	1,433

The **drug-protein** entity consists of drug-associated proteins. Considering that some clinical tests are conducted on mice and rats, drug-proteins for organisms including “mice” and “rats” in addition to the “Humans and other mammals” were extracted and imported into DDGICat. Table 3 summarizes the statistics of this entity.

Table 3: Drug-protein statistics a) Grouped by drug-protein b) Grouped by drug-protein and drug type

a)		b)		
Protein Type	#Records	Protein Type	#Small Molecule Drugs	#Biotech Drugs
Target	14,514	Target	13,623	891
Enzyme	5,179	Enzyme	5,027	152
Carrier	816	Carrier	791	25
Transporter	3,079	Transporter	3,050	29

The **DDI** entity has a total of 2.682,157 interacting drug pairs with name and interaction description attributes. This entity primarily consists of theoretical DDI entries, most of which did not have clinical evidence since they were produced with software and algorithms.

The **Drug-SNP** module contains drug-associated SNP records, including both adverse and other interactions. Each record has attributes such as drug identifier, drug-protein type, drug-protein identifier, drug-protein name, gene name, and description.

3.1.2 PharmGKB

PharmGKB content is provided partially in specialized files. The data files used for this study are the Clinical Annotations and Relationships. Therefore, these files were downloaded, parsed, and imported into the local RDBMS [March 2021].

The Clinical Annotations file has summaries of associations between drugs and genetic variants. This module has information on chromosomes, genes, variants, drugs, and diseases. After filtering out the drugs that do not exist in DrugBank, the proper text parsing operations were conducted on gene and drug columns of the remaining drugs. Finally, 5,897 annotation records for 509 drugs, 859 genes, 2,457 variants, and 2,345 PubMed identifiers were imported into DDGICat.

The relationships file has information on name, type, evidence, description, pharmacokinetics (PK), pharmacodynamics (PD), PubMed identifier, and the association information on variant-drug, variant-disease, gene-drug, gene-disease, gene-gene, and drug-drug. Sixty clinically-tested DDI entries were extracted from this file and imported into DDGICat. Table 4 summarizes the content of Clinical Annotations and Relationships files.

Table 4: PharmGKB data files a) Clinical annotations b) Relationships

a)		b)	
Entity	#Records	Entity	#Records
Drug	637	variant-drug	6,065
Variant	2,900	variant-disease	4,120
Gene	1,007	gene-drug	5,720
PMID	5,205	gene-disease	3,548
Total	4,559	gene-gene	2,836
		drug-drug	60

3.1.3 Ensembl

Ensembl has comprehensive biological content, including genes, variants, and phenotypes. Apart from being an extensive resource, Ensembl holds several naming conventions which may vary between knowledge bases. For instance, while DrugBank

represents protein information with a Uniprot identifier, PharmGKB holds them with an Ensembl identifier. Therefore, in this study, Ensembl was used to link the knowledge bases with different naming conventions.

Gene and SNP information was extracted from Ensembl, with an R package named “biomaRt” [51]. Gene information for the drug-proteins in DDGICat was extracted and put into a list object. Subsequently, this list was used to query the corresponding genes from Ensembl. SNP data were queried in a similar logic. Therefore, 3,897 genes and 538,615 SNPs were thus extracted from Ensembl. Table 5 summarizes the bioMart parameters used to query gene and SNP data.

Table 5: biomaRt parameters used to query data from Ensembl.

Entity	Gene	SNP
BioMart	ensembl	snp
Dataset	hsapiens_gene_ensembl	hsapiens_snp
Filters	hgnc_symbol	chr_name,start_position, end_position
Attributes	ensembl_gene_id,ensembl_transcript_id, hgnc_symbol,description,uniprot_gn_id, uniprot_gn_symbol,chromosome_name, start_position,end_position	refsnp_id,refsnp_source, chr_name,chrom_start, chrom_end

3.1.4 DDI Severity

The severity level of interacting drug pairs describes the importance of interaction. DDI records in DDGICat do not contain severity information. Therefore, this information was obtained from different knowledge bases, including KEGG Drug and ONC High.

DrugBank has the identifiers of drugs that exist in other knowledge bases. Based on the KEGG drug identifiers of 1,600 drugs existing in DrugBank, interaction severity values of 26,074 records were updated with the data retrieved from KEGG’s Rest Service. The same operation was conducted for ONC High, and 946 more records were updated with severity information.

Table 6: Severity information of interacted drug pairs retrieved from KEGG Drug and ONC High (CI: Contraindication, P: Precaution)

Source	KEGG Drug	KEGG Drug	KEGG Drug	ONC High	=27,007
Severity	P	CI	CI, P	high	
Record Count	25,492	337	232	946	

A quantitative summary of the data obtained after the earlier extraction processes are presented in Table 7 below.

Table 7: Statistical summary of DDGICat entities

Entity Name	Count
Drug	13,914
Gene	3,897
SNP	2,093
Drug-Protein	22,122
Drug-SNP	5,897
DDI	1,154,667
DDI pairs sharing same drug-protein	583,277
Disease	67

3.2 Data Integration

This section explains data integration steps of this study. As mentioned before, drug data is retrieved from DrugBank. DDI data is a combination of DrugBank, PharmGKB, KEGG Drug, and ONC High. Drug-associated SNP information is gathered from both DrugBank and PharmGKB. Disease information is extracted from PharmGKB. Finally, Ensembl is the central knowledge base for SNP and Gene information. Figure 1 depicts the data flows used in data integration.

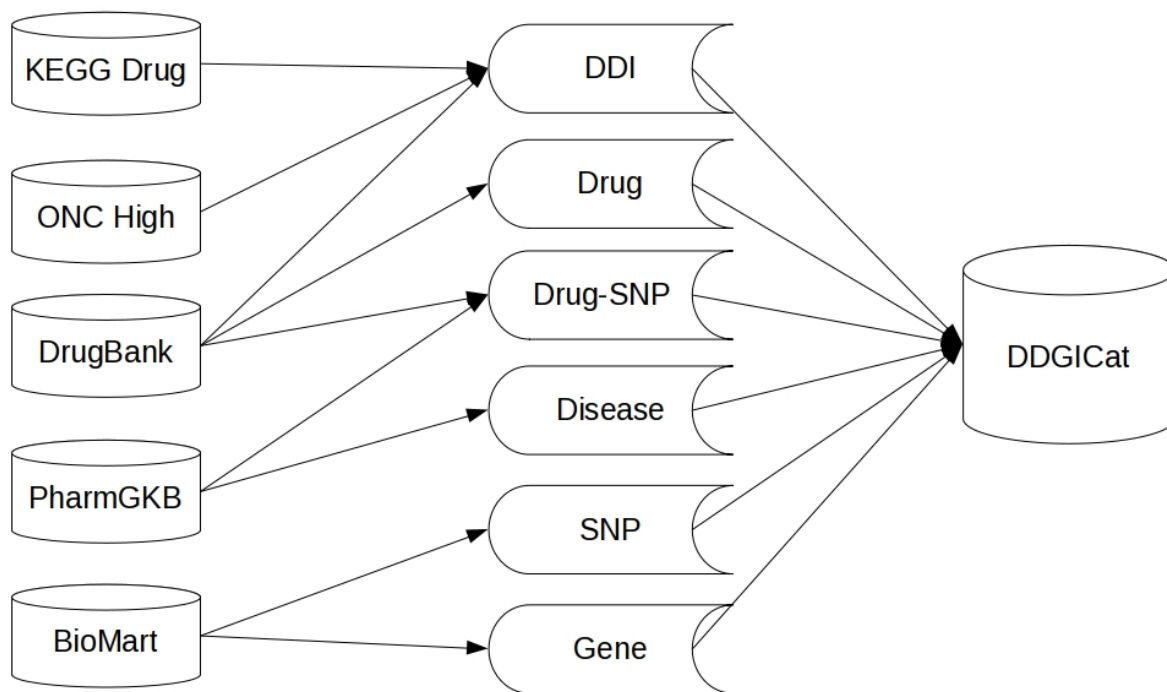


Figure 1: Data flows used in data integration

3.3 Data Preprocessing

This part summarizes the data preprocessing steps conducted for DrugBank and PharmGKB. In contrast, BioMart, KEGG Drug, and ONC High do not require data preprocessing since their content was already in the required format.

- DrugBank contains information on 222 different organisms, including bacteria, viruses, fungi, and mammals. Considering the use of mice or rats in the pre-experimental stages of drug development, groups other than “Human” and “Human and other mammals” were filtered out, and the remaining 13,185 drug records were imported into the DDGICat.
- Drug-associated proteins were combined into a single table with an additional attribute representing protein type (target, enzyme, carrier, and transporter).
- DDI records contained duplicate entries due to the directionality of drug pairs. Interacting drug pair combination was set as the primary key constraint. Additionally,

the lexicographical order of interacting drug pairs was used, and 1,154,667 distinct DDI records were determined out of 2,682,157 records.

- An additional interaction severity attribute was added to the DDI entity, and it was fed by the records retrieved from ONC High and KEGG Drug.
- Interacting drug pairs associated with the same drug-protein were calculated and stored in a new table named `ddi_same_drug_protein`.
- Drug-associated polymorphisms were retrieved from both DrugBank and PharmGKB and combined in a single table named `drug_snp`. Three hundred twenty-four entries were extracted from DrugBank. Five thousand nine hundred records were extracted from PharmGKB after conducting proper text parsing operations since “Gene” and “Drug Name” attributes were multi-valued (delimited by semicolons).

3.4 Database Creation

A relational database management system (RDBMS) enables the manipulation of stored data in a database. In this study, we used PostgreSQL 9.5.24 to store DDGICat content. PostgreSQL is an open-source RDBMS that dates back to 1986 as part of the POSTGRES project at the University of California [61].

DDGICat contains five entities, eight tables, and three connection tables. Details of them are as follows:

DRUG: This entity details information about drugs. We extract most of them from DrugBank. The attributes of this entity are as follows.

- `drug_id` (Drug identifier)
- `name` (Drug name)
- `synonym` (Drug synonym)
- `type` (Drug type)
- `description` (Drug description)
- `state` (Drug state)
- `indication` (Drug indication)
- `toxicity` (Drug toxicity)

- pharmacodynamics, absorption, half_life, metabolism, mechanism_of_action, volume_of_distribution, protein_binding, clearance, route_of_elimination (PK and PD) pubmed_id (PubMed identifier)

The primary key of this entity is drug_id.

DRUG PROTEIN: This entity contains drug proteins, including target, enzyme, carrier, and transporter. This entity has the following attributes.

- drug_id (Drug identifier)
- protein_id (Target/enzyme/carrier/transporter identifier)
- protein_type (Protein type)
- protein_name (Protein name)
- source (Protein source (TrEMBL, Swiss-Prot))
- uniprot_id (Uniprot identifier)
- gene_name
- function (A brief explanation of protein's functions)
- pubmed_id (PubMed identifier)

The (drug_id, protein_id, type) combination is the primary key of this entity.

GENE: Drug-related gene information extracted from DrugBank, PharmGKB, and Ensembl was combined and stored in this entity.

- ensembl_gene_id (Ensembl gene identifier)
- hgnc_symbol (Gene symbol)
- description
- uniprot_id (Uniprot identifier)
- uniprot_symbol (Uniprot symbol)
- chromosome (Chromosome name)
- start_position (Start position on the chromosome)
- end_position (End position on the chromosome)

The (ensembl_id, uniprot_id) combination is the primary key of this entity.

SNP: This entity holds SNP records. We extracted the required data from Ensembl/BioMart. This entity has the following attributes:

- refsnp_id (Single nucleotide polymorphism identifier)
- refsnp_source (Database source)
- chr_name (Chromosome name)
- chrom_start (Start position on the chromosome)
- chrom_end (End position on the chromosome)
- chrom_strand
- allele

refsnp_id attribute is the primary key of this entity.

DDI: This entity has adverse drug-drug interactions and potential drug-drug interactions. Its attributes are:

- drug1_id (Drug identifier of the first drug)
- drug2_id (Drug identifier of the second drug)
- description (A brief description of the interaction)
- category (Interaction category)
- drug1_name (Name of the first drug)
- drug2_name (Name of the second drug)
- severity (interaction severity level)
- severity_desc (interaction severity explanation)

The (drug1_id and drug2_id) set is the primary key of this entity.

DRUG_MAPPER: This is a connection table, which maps drug identifiers of different knowledge bases into each other. It has the following attributes:

- drugbank_id (Drugbank drug identifier)
- pharmgkb_id (PharmGKB drug identifier)
- onchigh_id (ONC High drug identifier)
- kegg_id (KEGG drug identifier)

The drugbank_id is the primary key of this table.

GENE-SNP: This is a mapper table and maps SNP records to the genes. Attributes of it are as follows:

- emsembl_id (Ensembl gene identifier)
- uniprot_id (Uniprot identifier)
- snp_id (Single nucleotide polymorphism identifier)

The (ensembl_id, uniprot_id, and snp_id) set is the primary key of this table.

DRUG-SNP: This is a connection table providing relations between drug and SNP entities. It has the following attributes:

- drug_id (Drug identifier)
- snp_id (Single nucleotide polymorphism identifier)
- uniprot_id (Uniprot identifier)
- gene_name
- chromosome
- significance
- description
- severity
- pubmed_id (PubMed identifier)

The primary key of this table is (drug_id, snp_id, uniprot_id, gene_name, chromosome, description, and pubmed_id) combination.

The entity-relationship diagram (ERD) is a logical model of the database, which depicts the attributes of each entity and the relations between them with specialized symbols. ERD of DDGICat is shown in Figure 2. Similarly, the Relational Data Model of DDGICat is depicted in Figure 3.



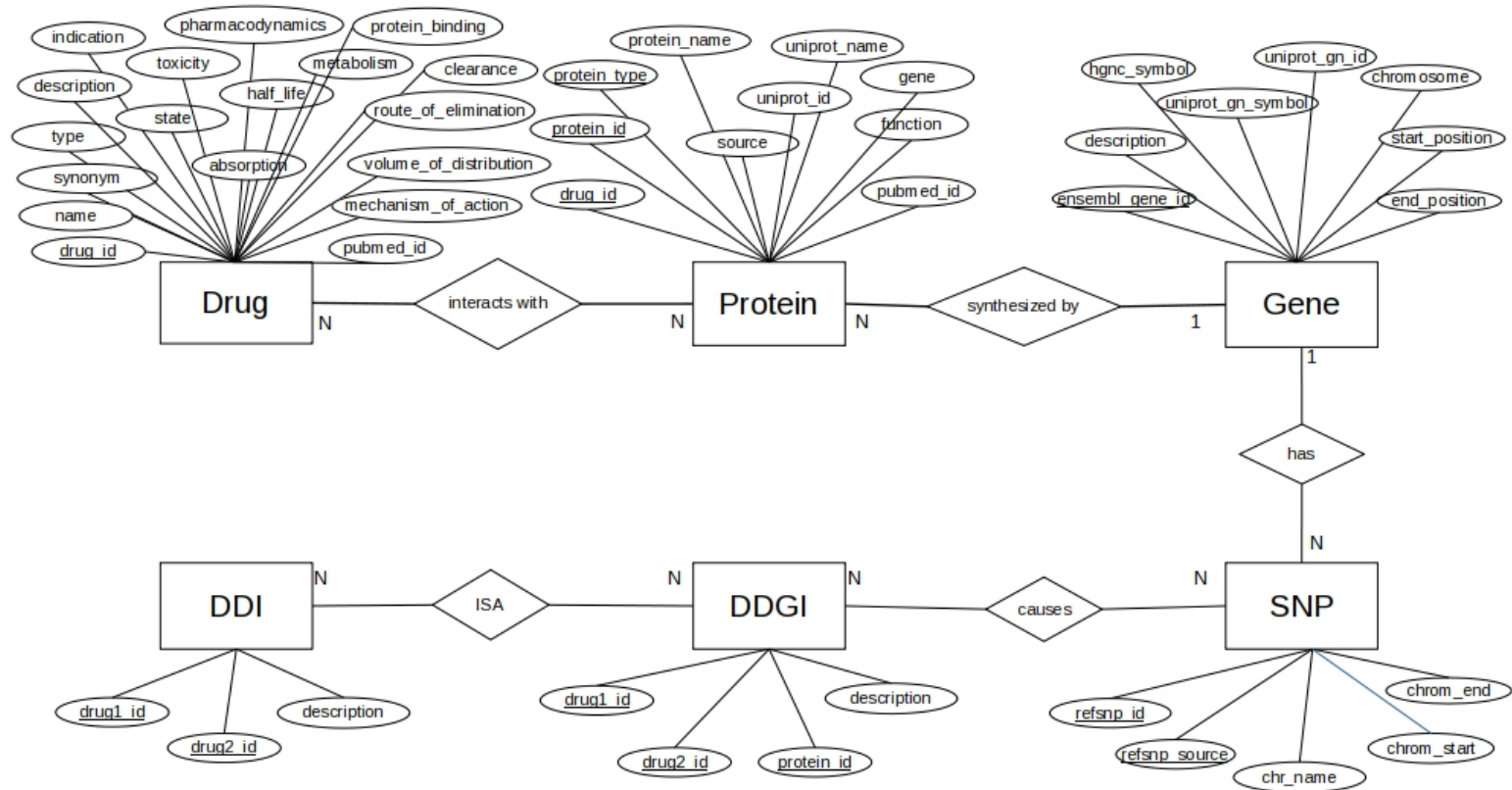


Figure 2: ER Diagram of DDGICat

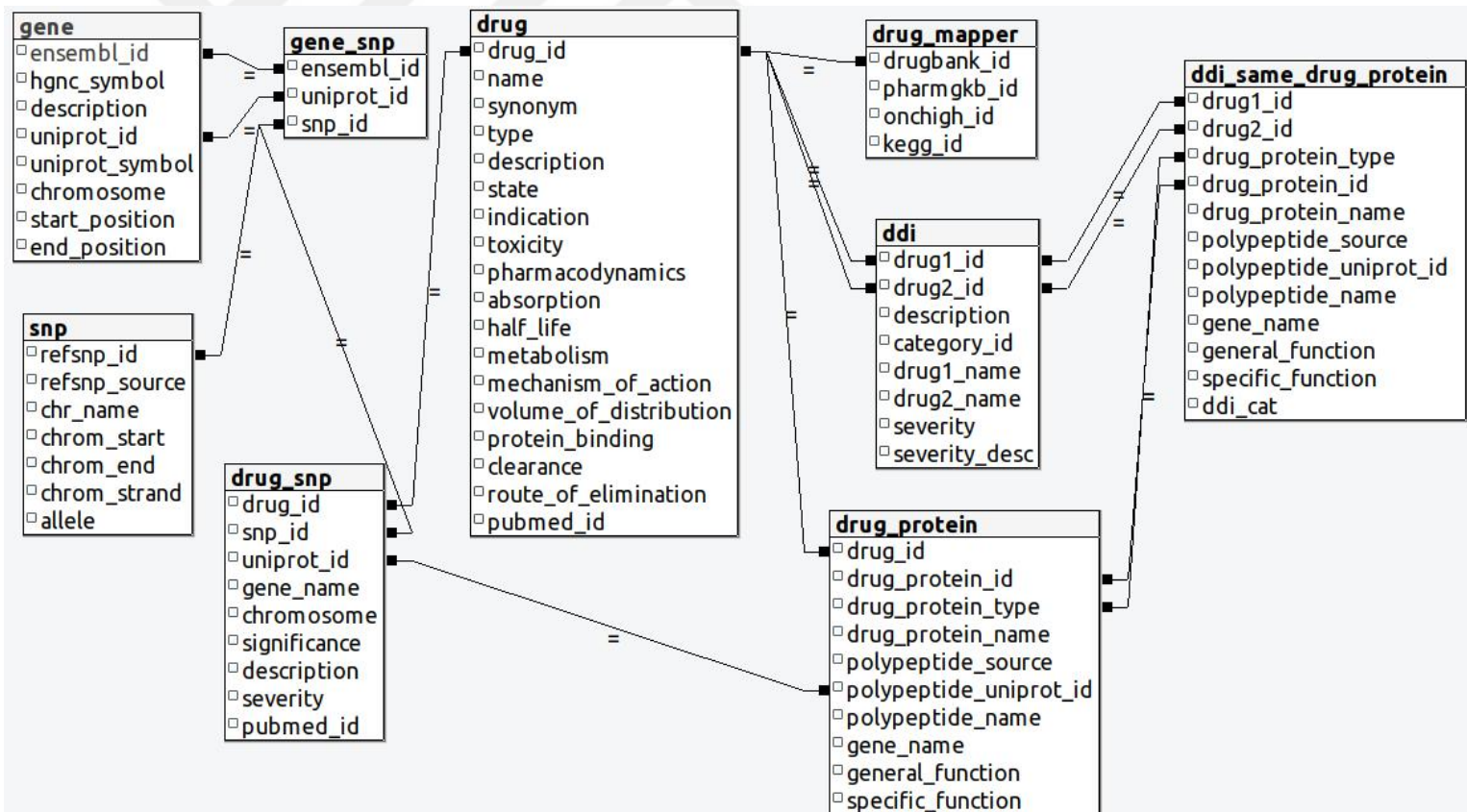


Figure 3: The Relational Data Model of DDGICat (Screenshot from PgAdmin 3)

3.5 Querying DDGICat

DDGICat is publicly accessible both by downloading the database content and a web application named DDGICat Browser. DDGICat Browser enables viewing DDGICat data in both tabular and graphical formats. DDGICat Browser was developed with an open-source R package named Shiny [62]. Shiny enables the development of interactive web applications. At the time of this study, the latest Shiny version was 1.6.0, published in January 2021 [62].

The Downloads page of DDGICat Browser provides the database entities as downloads, as shown in Figure 4.



Figure 4: Screenshot of Downloads Page on DDGICat Browser

Similarly, DDGICat Browser enables Drug, Gene, SNP, DDI, DGI, and DDGI contents to be viewed in tabular and graphical formats. The screenshots regarding these Pages are shown in Figure 5, Figure 6, Figure 7, and Figure 8, respectively. Additionally, the data retrieval logic of Drug, Gene, SNP, DDI, DGI, and DDGI Pages is depicted in Figure 9.

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Please enter drug name

Please select drug approval status

approved

Please select the drug type

☒ small molecule
 ☐ biotech

Copy

CSV

Excel

PDF

Print

Search:

Drug Id	Name	Type	State	Synonym
 DB00006	Bivalirudin	small molecule	solid	Bivalirudin

Indication For treatment of heparin-induced thrombocytopenia and for th...

Description Bivalirudin is a synthetic 20 residue peptide (thrombin inhi...

Toxicity Based on a study by Gleason et al., the no-observed-adverse-...

Pharmacodynamics Bivalirudin mediates an inhibitory action on thrombin by directly and specifically binding to both the catalytic s...

Absorption Following intravenous administration, bivalirudin exhibits linear pharmacokinetics. The mean steady state concentratio...

Half Life * Normal renal function: 25 min (in normal conditions) * Creatinine clearance 10-29mL/min: 57min * Dialysis-dependan...

Metabolism 80% proteolytic cleavage

Mechanism Of Action Inhibits the action of thrombin by binding both to its catalytic site and to its anion-binding exosite. Thrombin...

Volume Of Distribution 0.2L/kg

Protein Binding Other than thrombin and red blood cells, bivalirudin does not bind to plasma proteins.

Clearance

Route Of Elimination Bivalirudin is cleared from plasma by a combination of renal mechanisms (20%) and proteolytic cleavage.

Pubmed Id [16466327](#),[17381384](#),[16553503](#),[11156732](#),[21108549](#),[16614733](#),[12851152](#)

Figure 5: Screenshot of Drug Page on DDGICat Browser

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Please enter gene name

DCK

Copy

CSV

Excel

PDF

Print

Search:

Ensembl Id	Hgnc Symbol	Uniprot Id	Uniprot Symbol	Chromosome	Start Position	End Position	Description
ENSG00000156136	DCK	DCK	DCK	4	71858255	71896631	deoxycytidine kinase [Source:HGNC Symbol;Acc:2704]

Showing 1 to 1 of 1 entries

Previous

1

Next

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Please enter SNP name

Copy

CSV

Excel

PDF

Print

Search:

RefSNP Id	RefSNP Source	Chromosome Name	Chromosome Start	Chromosome End	Chromosome Strand	Allele
rs2301159	dbSNP	13	103045378	103045378	1	G/A
rs222749	dbSNP	17	3592080	3592080	1	G/A
rs588765	dbSNP	15	78573083	78573083	1	T/A/C/G
rs16973225	dbSNP	15	81937658	81937658	1	A/C

Figure 6: Screenshot of Gene and SNP Pages on DDGICat Browser

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Please enter Drug1 name

Quinidine

Please enter Drug2 name

warfarin

Please select severity level

all

Copy

CSV

Excel

PDF

Print

Search:

Drug1 Id	Drug2 Id	Drug1 Name	Drug2 Name	Severity	Description
DB00682	DB00908	Warfarin	Quinidine		The serum concentration of Warfarin can be increased when it is combined with Quinidine.

Showing 1 to 1 of 1 entries

Previous

1

Next

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Please Enter Drug Name

Please Select Drug Protein Type

all

Please Enter Gene Name

Copy

CSV

Excel

PDF

Print

Search:

Drug Id	Name	Protein Type	Gene Name	Drug Protein Name	Polypeptide Uniprot Id	General Function
+ DB00001	Lepirudin	Target	F2	Prothrombin	P00734	Thrombospondin receptor activity
+ DB00002	Cetuximab	Target	FCGR3B	Low affinity immunoglobulin gamma Fc region receptor III-B	O75015	
+ DB00002	Cetuximab	Target	C1QA	Complement C1q subcomponent subunit A	P02745	

Figure 7: Screenshot of DDI and DGI Pages on DDGICat Browser

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Plese select a disease

Depressive Disorder

Plese select drug 1

escitalopram

Plese select drug 2

ethanol

Shared

☐ Chromosome
 ☒ Gene
 ☐ Protein
 ☐ SNP

Copy

CSV

Excel

PDF

Print

Search:

Gene Name	Drug1 Name	Drug2 Name	Drug Protein Type	Drug Protein Name	Uniprot Id	Polypeptide Name	General Function
 CYP3A4	Escitalopram	Ethanol	Enzyme	Cytochrome P450 3A4	P08684	Cytochrome P450 3A4	Vitamin d3 25-hydroxylase activity
Specific Function Cytochromes P450 are a group of heme-thiolate monooxygenases. In liver microsomes, this enzyme is involved in an NADPH-dependent electron transport pathway. It performs a variety of oxidation reactions (e.g. caffeine 8-oxidation, omeprazole sulfoxidation, midazolam 1'-hydroxylation and midazolam 4-hydroxylation) of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. Acts as a 1,8-cineole 2-exo-monooxygenase. The enzyme also hydroxylates etoposide (PubMed:11159812). Catalyzes 4-beta-hydroxylation of cholesterol. May catalyze 25-hydroxylation of cholesterol in vitro (PubMed:21576599).							
 CYP2C19	Escitalopram	Ethanol	Enzyme	Cytochrome P450 2C19	P33261	Cytochrome P450 2C19	Steroid hydroxylase activity

Showing 1 to 2 of 2 entries

Previous

1

Next

Figure 8: Screenshot of DDGI Page on DDGICat Browser. The DDGI Page shows the relation between a disease and a drug pair in SNP, protein, gene, and chromosome detail

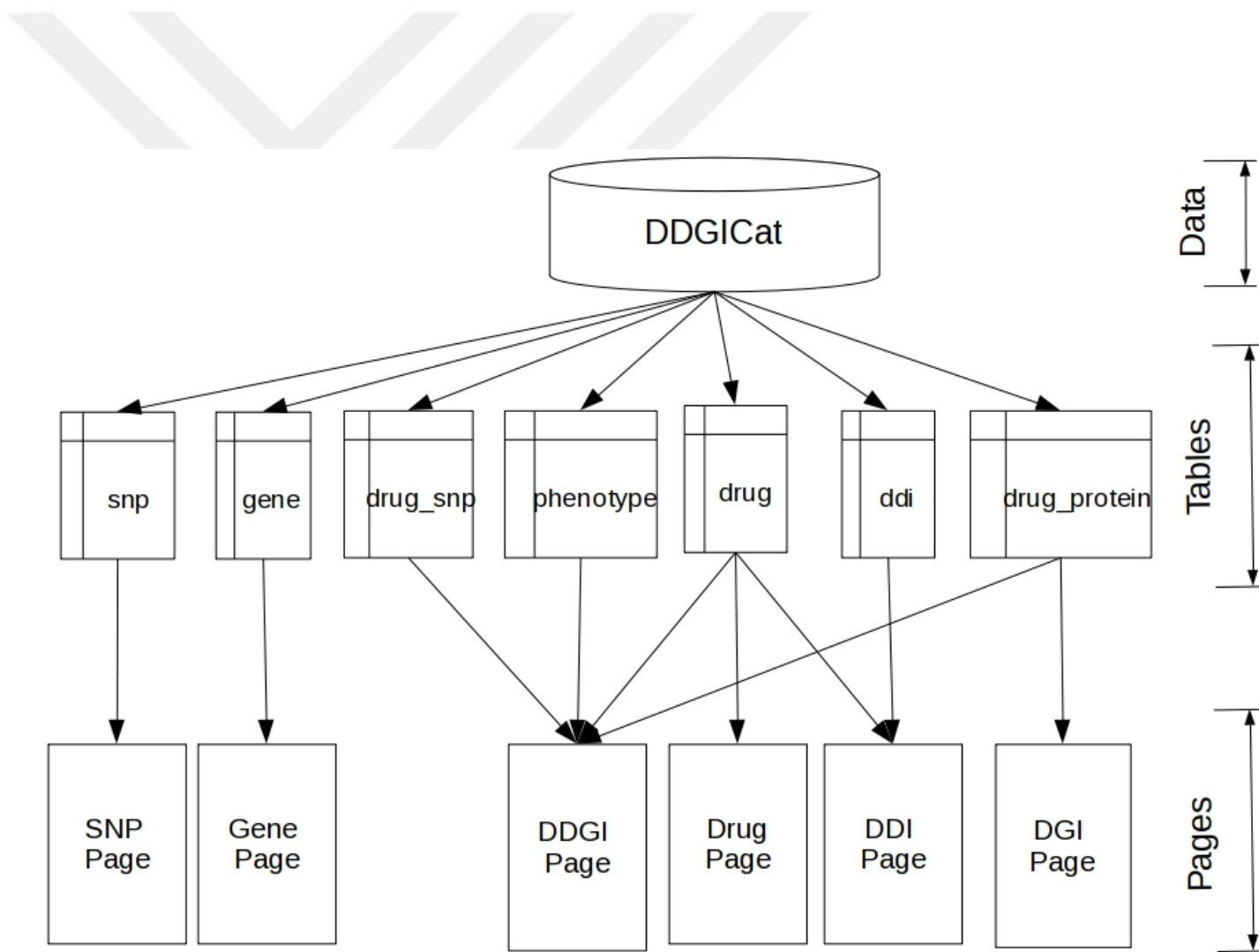


Figure 9: Data retrieval logic behind Drug, Gene, SNP, DDI, DGI, and DDGI Pages.



CHAPTER 4

RESULTS

DDGICat content, which consists of relationships between diseases, drugs, genes, drug proteins, and SNPs, aims to guide different consumer types such as researchers, prescribers, pharmacists, and statisticians. This chapter provides sample usages of DDGICat in different scenarios. Moreover, through the end of this chapter, examples of graphical outputs retrieved from DDGICat are shared.

4.1 Case Study

The DDI Page of DDGICat Browser displays the details of interacting drug pairs. Therefore, this page aims to analyze the possibility of whether drug pairs may lead to unpredicted DDIs. In the scenario to be exemplified, Acute Coronary Syndrome (ACS) was selected. ACS is a general term used to describe the conditions associated with the decrease in the blood flow to the heart. One of the leading causes of ACS-associated diseases is genetics.

Clopidogrel is a drug used to treat ACS-related diseases. **DDI** Page provides possible drug interactions with Clopidogrel. Additionally, the interaction severity filter enables more focused searches on critical interactions.

DGI Page displays Clopidogrel-associated proteins (target, enzyme, carrier, transporter). This page also allows users to filter the result set for a specific drug-associated protein or gene. Figure 10 demonstrates the screenshots of DDI and DGI Pages.

DDGI Page presents the cumulative effect of DDI and DGI. For instance, similar to the previous scenario, in the case of a patient, who uses Clopidogrel for her heart disease, if the patient adds a second drug such as Omeprazole to treat her Gastroesophageal Reflux Disease (GERD), these two drugs may interact.

As shown in Figure 11, DDGI Page allows prescribers to examine the interaction of Clopidogrel and Omeprazole in detail in terms of chromosomes, genes, proteins, and

SNPs. According to the results retrieved from this page, drug-proteins associated with Clopidogrel and Omeprazole stand on chromosomes 7, 10, 15, and 19. In addition, “P-glycoprotein 1” synthesized by the ABCB1 gene is the carrier protein for Clopidogrel and Omeprazole. Similarly, enzymes including CYP1A2, CYP3A4, CYP2C9, and CYP2C8 metabolize these two drugs. Moreover, the shared polymorphism for Clopidogrel and Omeprazole is rs104564.



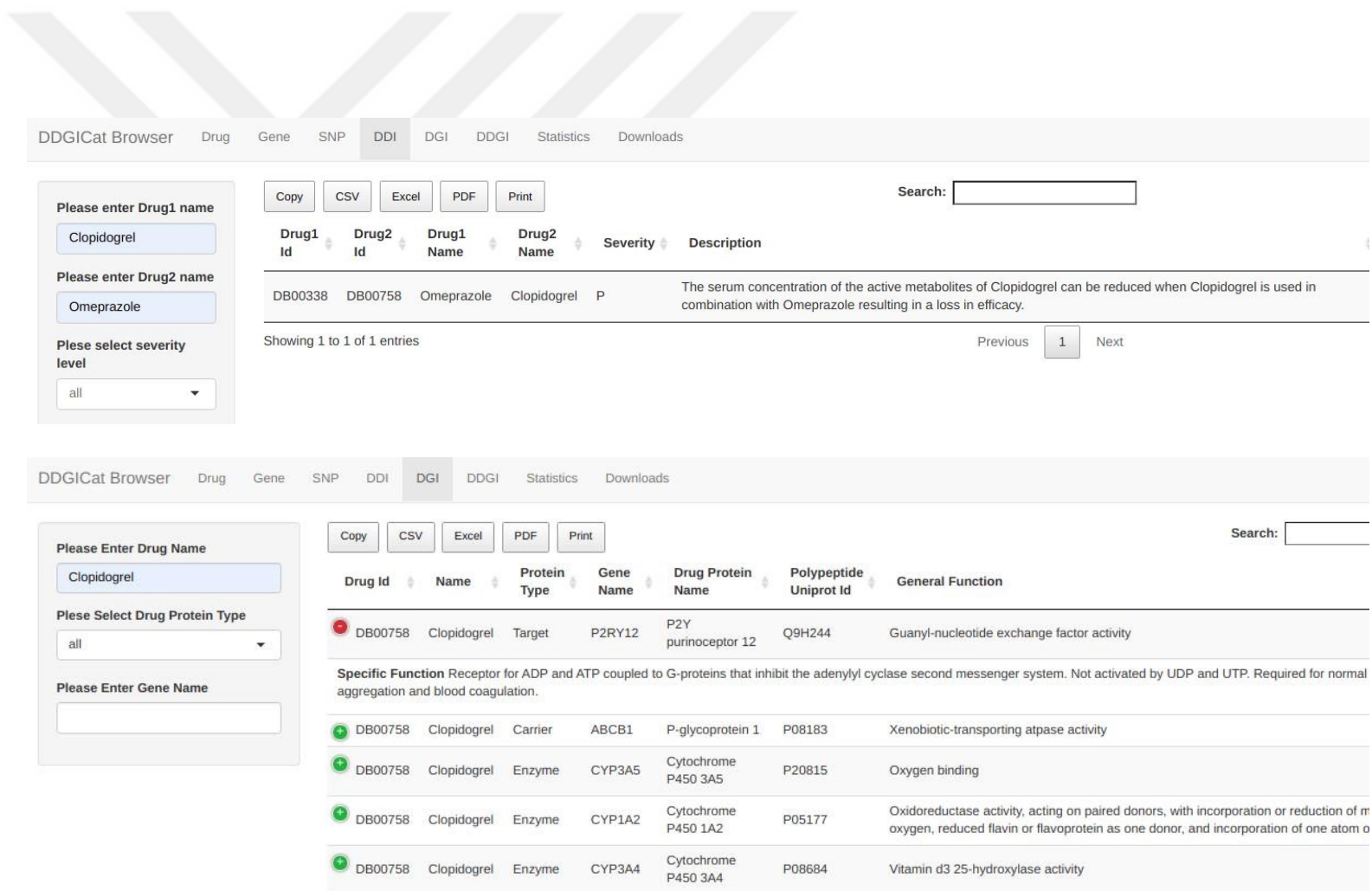


Figure 10: Drugs and drug-associated proteins interacting with Clopidogrel. (Screenshot of DDI and DGI Pages)

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Plese select a disease

Acute coronary syndrome

Plese select drug 1

clopidogrel

Plese select drug 2

omeprazole

Shared

☐ Chromosome
 ☒ Gene
 ☐ Protein
 ☐ SNP

Copy CSV Excel PDF Print

Search:

Gene Name	Drug1 Name	Drug2 Name	Drug Protein Type	Drug Protein Name	Uniprot Id	Polypeptide Name	General Function
ABCB1	Clopidogrel	Omeprazole	Carrier	P-glycoprotein 1	P08183	Multidrug resistance protein 1	Xenobiotic-transporting atpase activity
Specific Function Energy-dependent efflux pump responsible for decreased drug accumulation in multidrug-resistant cells.							
CYP1A2	Clopidogrel	Omeprazole	Enzyme	Cytochrome P450 1A2	P05177	Cytochrome P450 1A2	Oxidoreductase activity, acting on paired donor with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen
CYP3A4	Clopidogrel	Omeprazole	Enzyme	Cytochrome P450 3A4	P08684	Cytochrome P450 3A4	Vitamin d3 25-hydroxylase activity
				Cytochrome P450		Cytochrome P450	

DDGICat Browser

Drug

Gene

SNP

DDI

DGI

DDGI

Statistics

Downloads

Plese select a disease

Acute coronary syndrome

Plese select drug 1

clopidogrel

Plese select drug 2

omeprazole

Shared

☐ Chromosome
 ☐ Gene
 ☐ Protein
 ☒ SNP

Copy CSV Excel PDF Print

Search:

SNP Name	Chromosome	Gene Name	Drug1 Name
rs1045642	chr7	ABCB1	Clopidogrel
Drug2 Name Omeprazole			
rs1045642	chr7	ABCB1	Clopidogrel
rs1045642	chr7	ABCB1	Clopidogrel
rs1045642	chr7	ABCB1	Clopidogrel
rs1045642	chr7	ABCB1	Clopidogrel
rs1045642	chr7	ABCB1	Clopidogrel

Showing 1 to 6 of 6 entries

Previous

1

Next

Figure 11: Severity of Clopidogrel and Omeprazole interaction affected by genetic materials. (Screenshots of DDGI Page)

4.2 Database Statistics

This section provides graphical outputs retrieved from DDGICat entities, including drug, gene, SNP, DDI, DGI, and DDGI. These graphs and more could be viewed on the Statistics Page of DDGICat Browser.

The **drug** entity contains information for two drug types: biotech and small-molecule. According to the FDA approval status, seven drug statuses include approved, experimental, illicit, investigational, nutraceutical, vet-approved, and withdrawn. Figure 12 provides a summary of the drug entity according to the mentioned categorizations. As shown in Figure 12, small molecule drugs form the vast majority of records. The experimental, investigational, and approved drug statuses have the highest proportions, respectively.

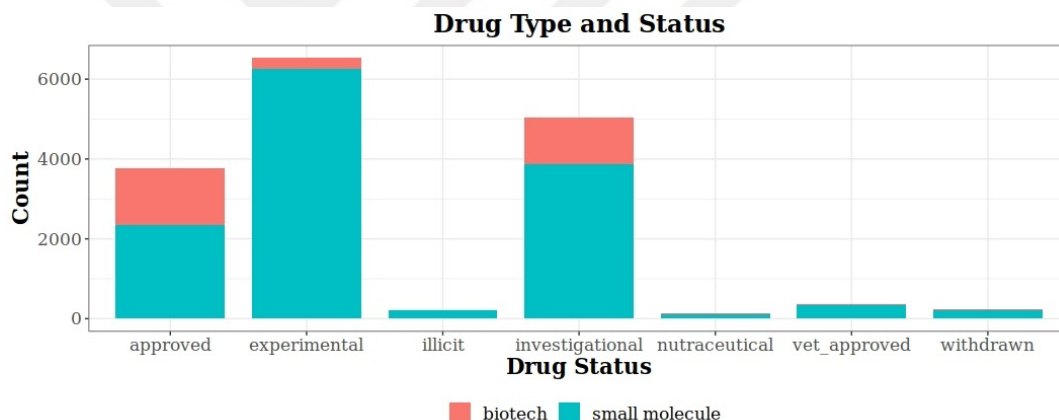


Figure 12: Drug types classified per drug status

In addition, the **gene** entity contains 3,306 distinct genes distributed on 17 chromosomes. The first and second chromosomes contain the highest number of genes, with 343 and 233 genes. Figure 13 demonstrates the chromosome distribution of drug-associated genes prepared based on the DDGICat and Ensembl databases.

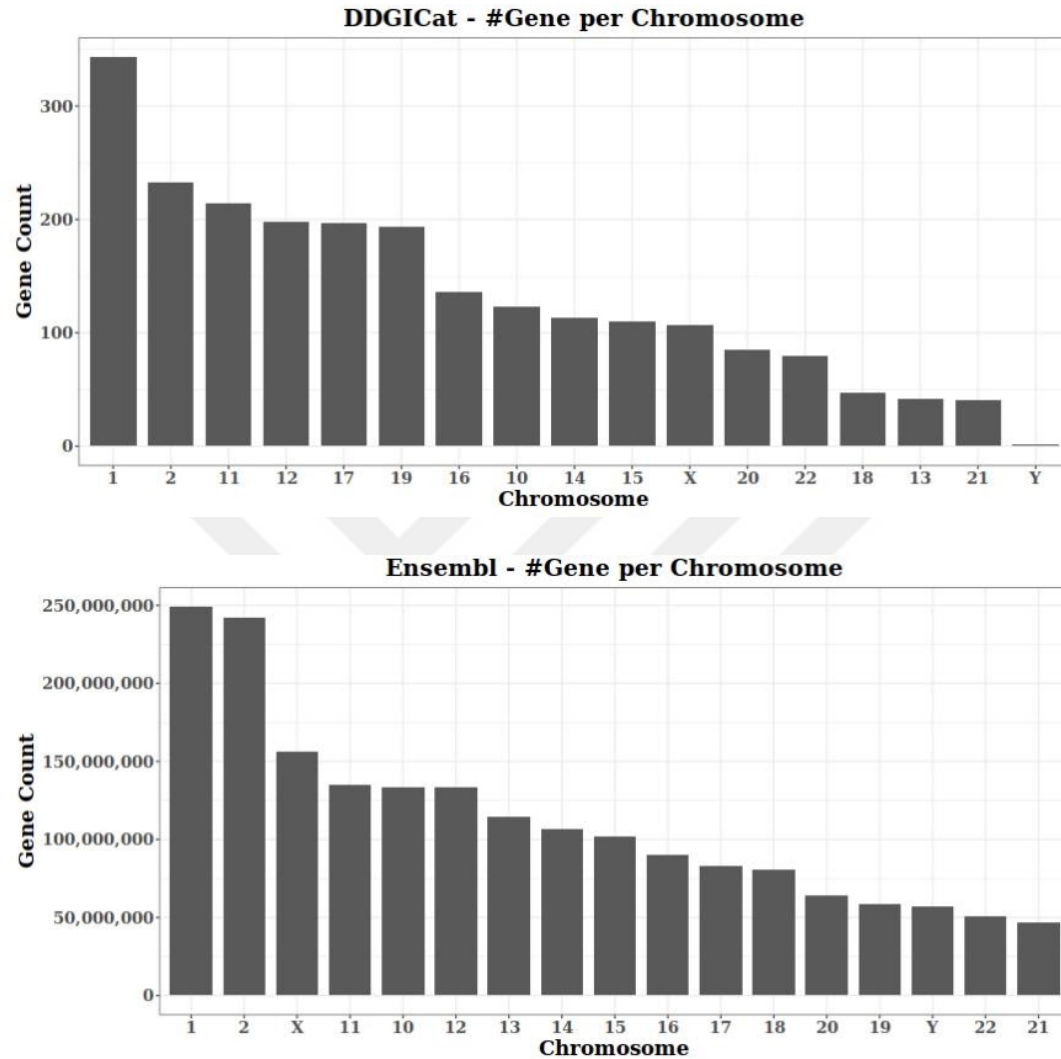


Figure 13: Drug-related gene distribution on chromosomes

SNP entity contains 2,093 drug-associated polymorphisms distributed on 22 chromosomes. As shown in Figure 14, the first chromosome has the most drug interacting polymorphic genes with 225 records. The exact figure also contains the same distribution based on the data retrieved from Ensembl.

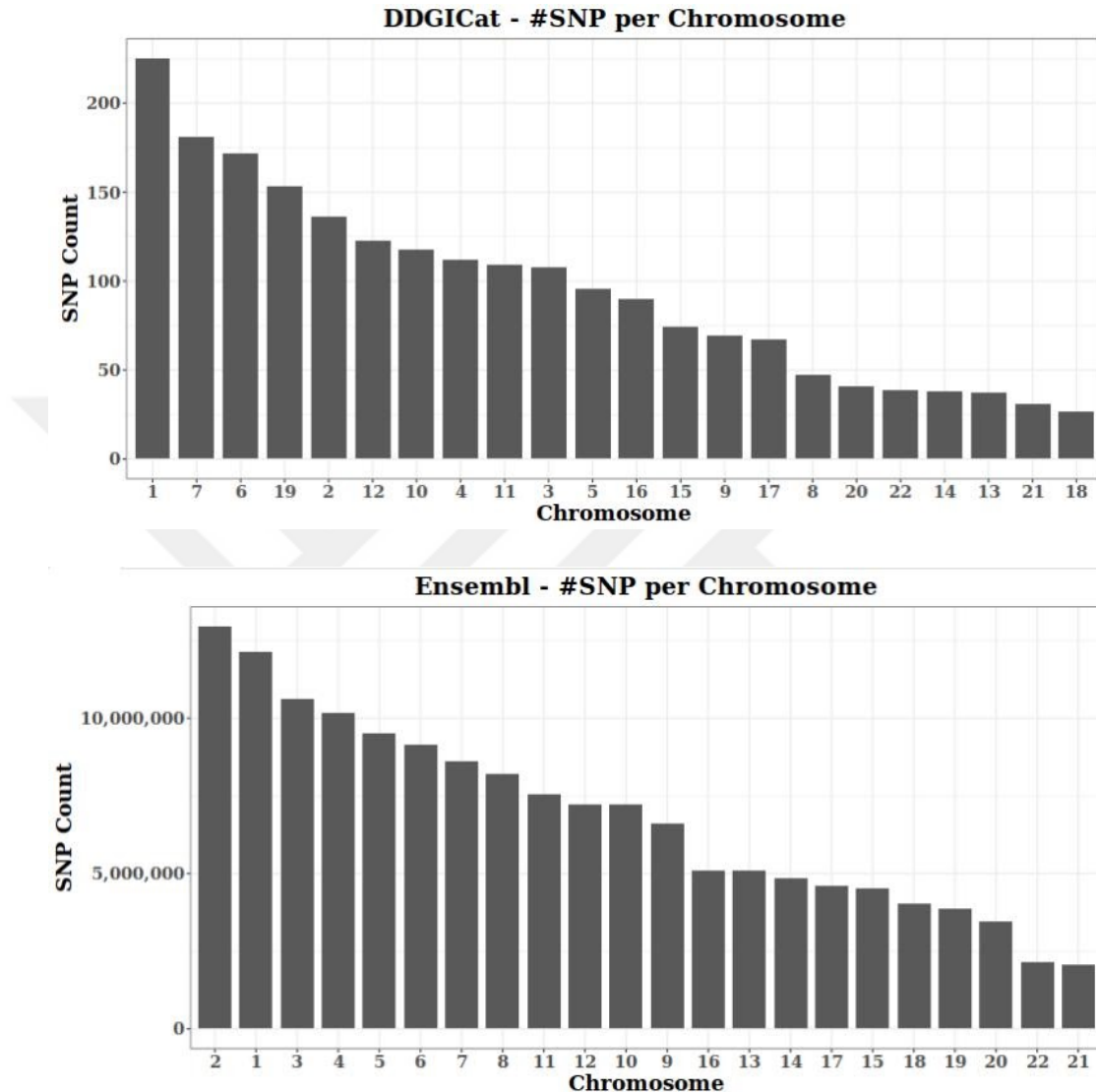


Figure 14: Distribution of drug-related SNPs on chromosomes

Additionally, Figure 15 depicts the top 10 most drug-associated genes. CYP3A4 (Cytochrome P450 3A4) is the gene interacting with most drugs (880 drugs).

The **drug-protein** entity holds information of drug-associated proteins, which are gene products that interact with drugs. There are 3,190 distinct drug-associated proteins, 2,906 of which are targets, 382 of which are enzymes, 77 of which are carriers, and 258 of which are transporters. As shown in Figure 16, the most drug-associated

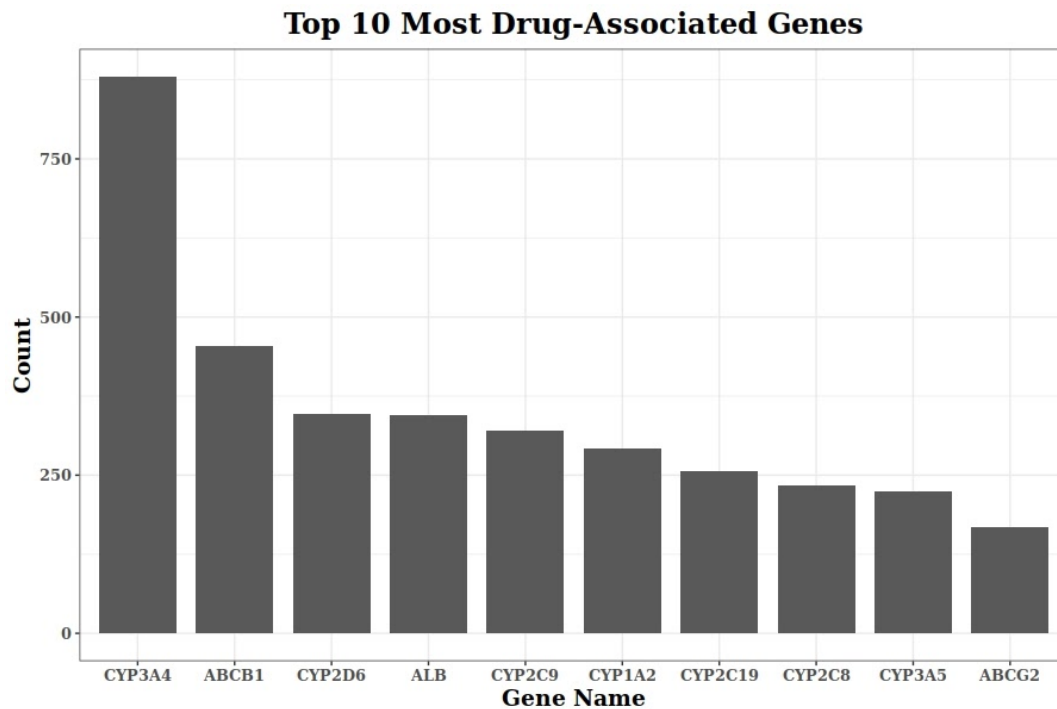


Figure 15: Top 10 genes interacting with most drugs

protein types are targets, enzymes, transporters, and carriers, respectively. Drug protein types with the highest potential to interact with drugs are Enzymes with 67%, Transporters with 16%, Targets with 13%, and 5% Carriers.

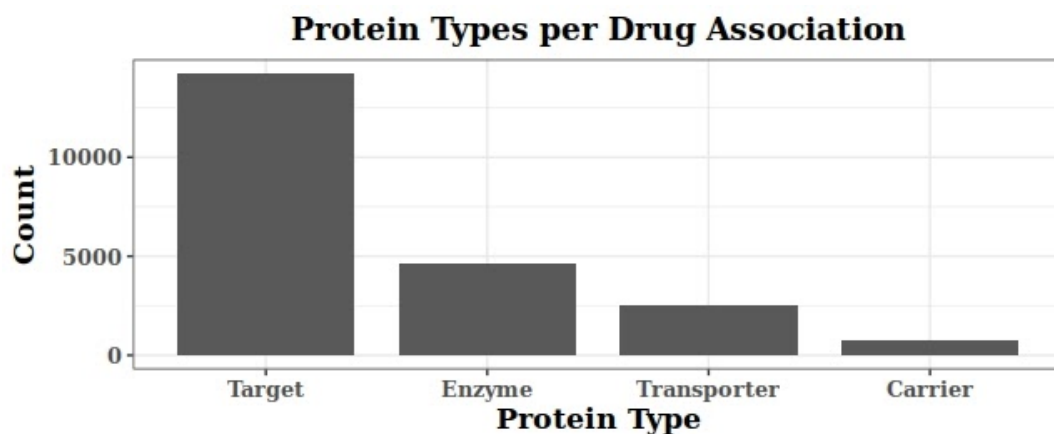


Figure 16: Distribution of drug-interacting protein types

As summarized in Figure 17, CYP3A4 (Cytochrome P450 3A4), the enzyme responsible for the metabolism of about 50% of drugs, has the highest drug interaction potential.

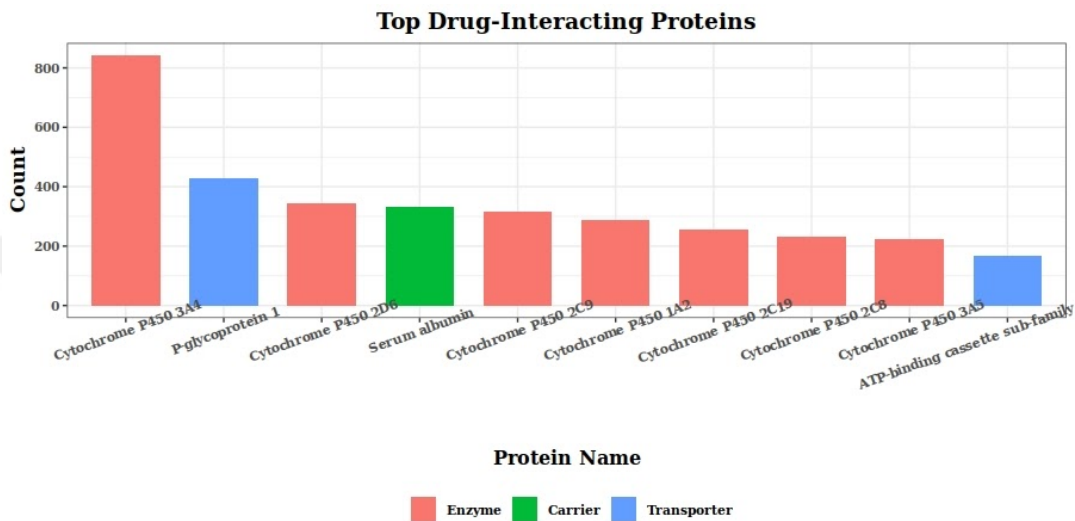


Figure 17: Top 10 drug-interacting proteins

Figure 18 summarizes the drug-protein distribution per drug. Most drugs (3,569) have a single target. Similarly, most drugs (670) are metabolized by a single Enzyme. Furthermore, most drugs have one or two carriers. In addition, the vast majority of drugs have one to three carriers and transporters.

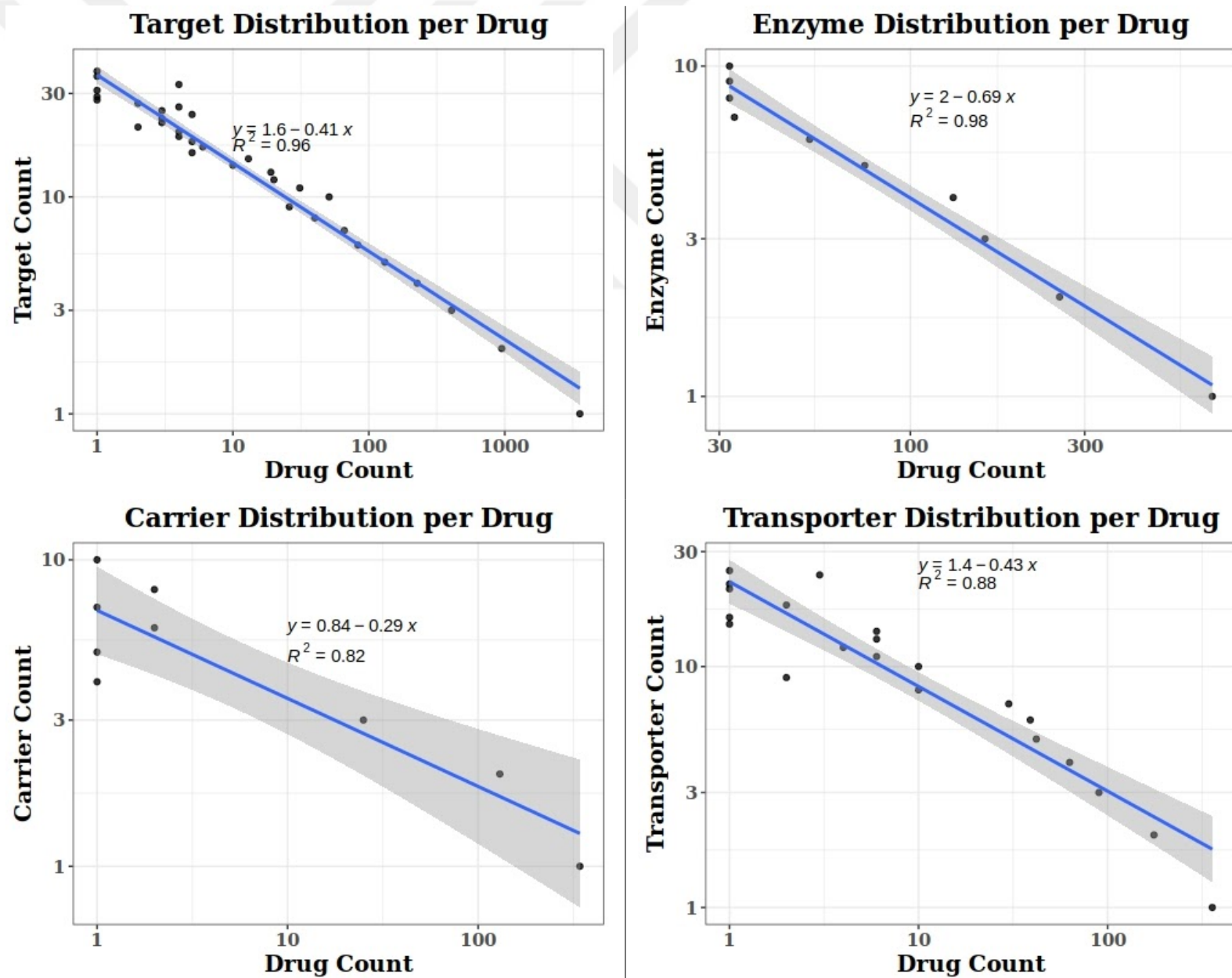


Figure 18: Drug-protein distribution per drug

The **DDI Entity** has more than one million distinct drug-drug interaction records on 3,952 drugs. These records, which constitute the majority of the content of this table, are mainly generated by using a prediction system in combination with drug labels and scientific publications. In the analyses related to drug-drug interaction entries, in order to get more significant results, records having high severity levels were included in the analysis in the following parts. For instance, DDI Distribution per drug is depicted in Figure 19. According to the figure, drug pairs mainly interact with one or two drugs. Additionally, the number of drugs that each drug interacts with varies between 1 and 40.

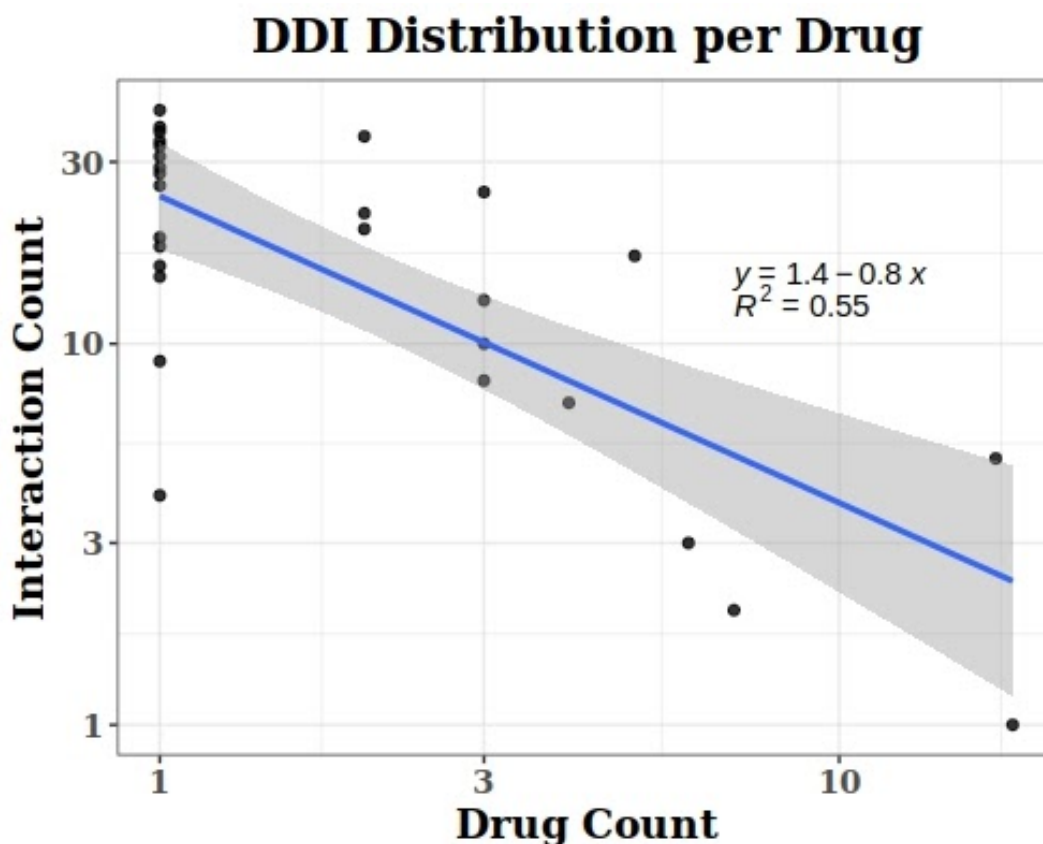


Figure 19: DDI distribution per drug

DDGICat has information on interacting drug pairs which share the same drug-protein, in total 583,277 records, 73,690 of which are targets, 392,264 of which are enzymes, 26,839 of which are carriers, and 90,484 of which are transporters. Figure 20 depicts the distribution of interacting drug pairs per having the same drug protein.

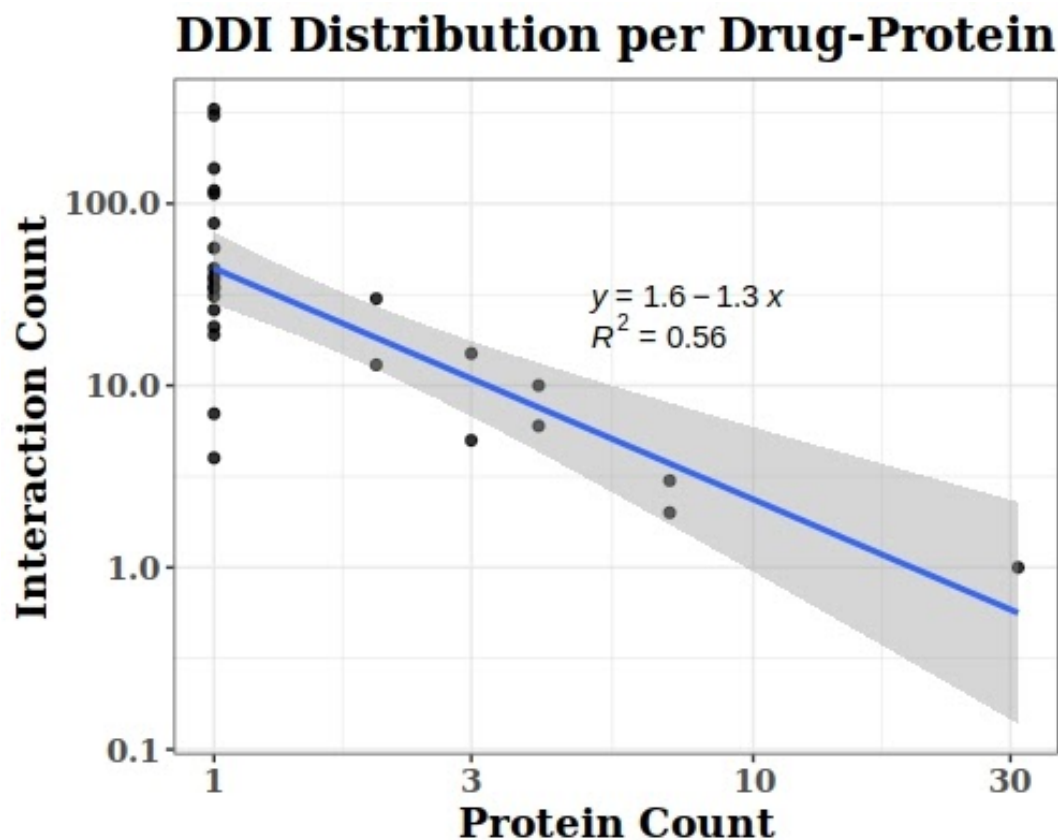


Figure 20: DDI distribution per drug-protein

Additionally, Figure 21 shows interacting drug pair percentages that share the same drug protein. According to the figure, drugs metabolized by the same enzyme interact at a rate of 64%. Similarly, drugs that share the same carrier and transporter interact at a rate of 70% and 56%, respectively. Moreover, drugs that share the same target interact at a rate of 6%.

Moreover, we searched if there was a correlation between interacting drug pairs and their ATC level. There are four ATC levels, and the first ATC level is called the root level. Figure 22 shows the drug-drug interaction distribution per ATC level of interacting drug pairs. As can be seen from the graph, the interaction rates of drugs increase as the ATC Level increases.

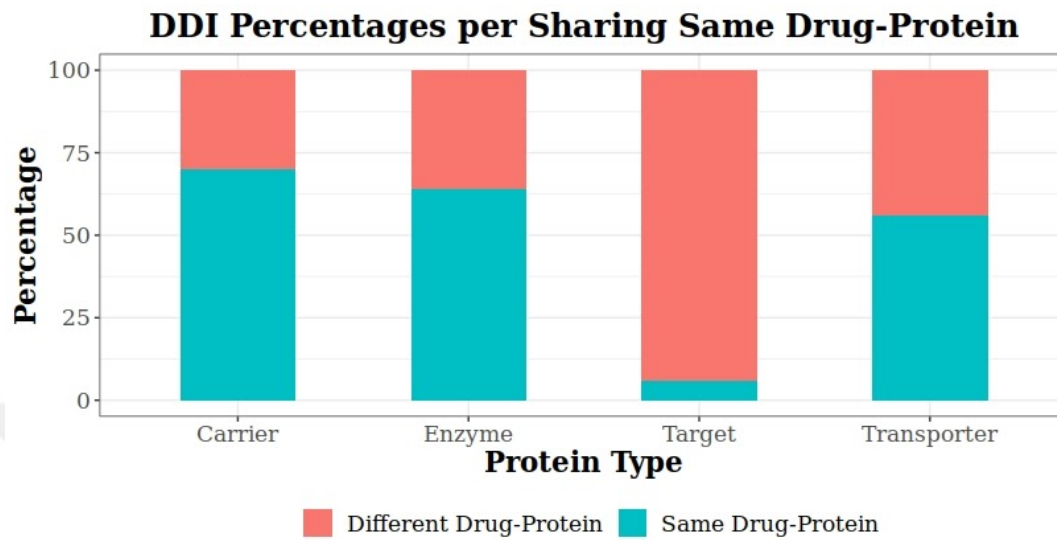


Figure 21: DDI Percentages per drug-protein

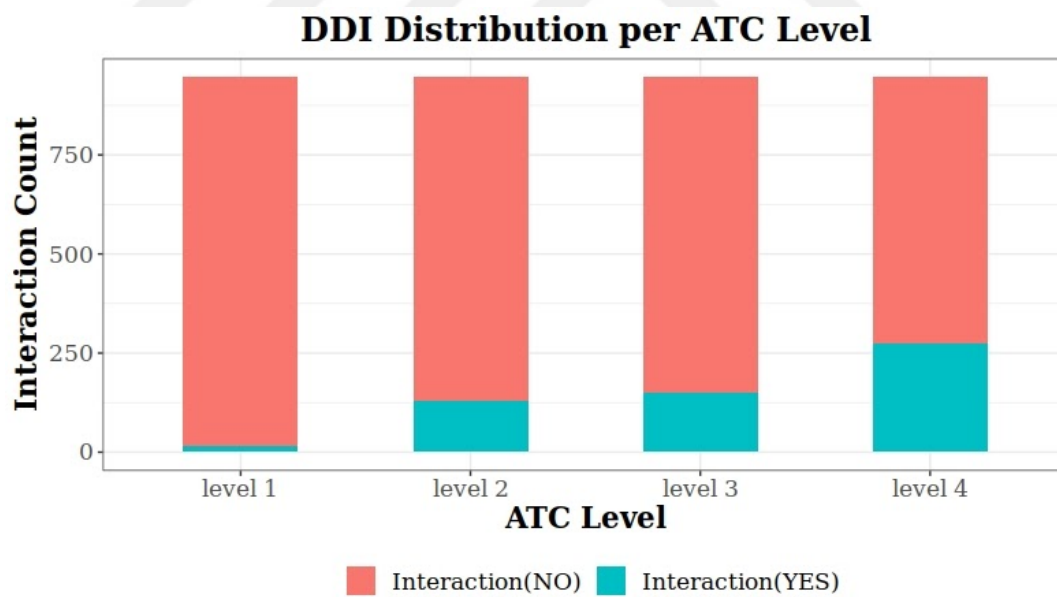


Figure 22: DDI distribution per ATC level

Furthermore, we classified interacting drug pairs into five main groups based on the description information of interacting drug pairs. Classification results are shown in Table 8.

Table 8: Interacting drug pair classification based on interaction description information (The descriptions are taken verbatim from DrugBank)

Id	Interaction Description	%
1	“The risk or severity of adverse effects can be increased when Drug A is combined with Drug B.”	49
2	“Drug A may increase/decrease the . . . activities of Drug B.”	19
3	“The metabolism of Drug A can be increased/decreased when combined with Drug B.”	16
4	“The therapeutic efficacy of Drug A can be decreased when used in combination with Drug B.”	12
5	“The serum concentration of Drug A can be increased/decreased when it is combined with Drug B.”	4

CHAPTER 5

DISCUSSION & CONCLUSION

5.1 Importance of Drug-Drug Interactions

Polypharmacy, especially for the elderly, plays a vital role in today's therapeutics. Although polypharmacy is commonly preferred in treatment, it may cause negative consequences. Drug-drug interactions (DDIs) are one of the leading drawbacks of polypharmacy. A significant amount of emergency room visits, hospital admissions, and rehospitalizations are due to DDIs.

DDIs threaten human health and pose an economic burden, necessitates a better understanding of their underlying reasons. Undoubtedly, one of the essential materials contributing to this challenge is DDI data. Therefore, the primary motivation of this study is to contribute to the ongoing DDI studies by creating a catalog database so that the required DDI data could be reached and used by researchers, prescribers, and the pharmaceutical industry.

5.2 Importance of Drug-Drug-Gene Interactions

Drug-drug-gene interaction (DDGI) is the cumulative effect of drug-drug interaction (DDI) and drug-gene interaction (DGI). DDGI occurs when genetic variations alter the concomitantly used drug pairs' pharmacokinetics (PK) or pharmacodynamics (PD).

Today, it is an accepted fact that drug response may vary between individuals. Among several triggering factors, such as age, gender, weight, and other medications, genetic variability is one of the most effective ones for these reasons. Therefore, the "one size fits all" approach has started to lose its sufficiency, and this approach is being replaced by Pharmacogenomics (PGx), a subfield of pharmacology that studies how genes cause altered drug responses.

Ongoing PGx studies have also contributed to DDI studies, and DDI researchers have enlarged their vision by adding genetics into the equation, and they are starting to inspect DDIs from the perspective of genetics.

5.3 Importance of DDGICat

To the best of our knowledge, DDGICat is the first catalog that allows drug-drug interaction (DDI) information to be searched and viewed together with genetic makeup, specifically at the chromosome, gene, protein, and SNP detail. Before this study, as far as we know, one other study (DDIPred, mentioned in Chapter 2) has also contributed to the gap in the ongoing DDI and genetic variability relationship. However, unlike previous work, DDGICat provides a combined dataset including DDI and drug-gene interaction (DGI) records and their associations. DDGICat broadens the content of genotype focussed drug-drug interaction data by having more entries than previous studies. Furthermore, different from existing studies, interacting drug pairs or drug-gene pairs have the option to be viewed together with shared chromosome, gene, protein, and SNP details. The content of DDGICat could be reached through the developed web portal, named DDGICat Browser.

5.4 Significance of the Data in DDGICat

In the previous section, the distribution of drugs based on their types was shared in Figure 12. According to this distribution, 17% of the data consists of biotech drugs, while 83% are small molecule drugs. Most drugs are composed of small-molecule drugs with a much longer history than biotech drugs, which is an understandable explanation for their higher quantity.

According to the Ensembl data, the first and second chromosomes have the highest number of genes, respectively (Figure 13). The same figure also shows the drug-associated gene distribution on the chromosomes in DDGICat data. Similar to the Ensembl data, the chromosomes with the highest number of genes are the first and second chromosomes, respectively. Therefore, considering the first two chromosomes, the number of genes in these chromosomes and the number of drug interacting genes are consistent.

Figure 15, and Figure 17 show the drug-interacting proteins. Based on these figures, proteins having the most drug interactions are CYP3A4, CYP2D6, CYP2C9, CYP1A2, and CYP2C19, respectively. In other words, based on the DDGICat data, most of the drug-associated proteins belong to the CYP 450 family. As is known, en-

zymes belonging to the CYP 450 family have a large proportion in drug metabolism. Therefore, this fact is confirmed by these results obtained from DDGICat.

As described in Chapter 2, the main aim in the drug development process is to determine the target protein causing the disease and fix or mitigate the health problem by providing the conditions in which the appropriate drug dosage will act on the target protein. Considering this information, the distribution in Figure 16 is meaningful; since the target protein quantity is much higher than other drug proteins, including enzyme, transporter, and carrier.

In the previous section, Figure 18 summarizes the drug-protein (target, enzyme, carrier, transporter) distribution per drug. This relationship follows a power-law distribution, indicating that the network of drug-protein interactions is a scale-free network. This scale-free nature exists both as a whole and at the level of different protein classes (target, enzyme, carrier, transporter). That, in turn, indicates that there are “hub” drugs, as well as hub proteins (like the aforementioned CYP450 family).

The previous section shared the analysis of drug-drug interaction (DDI) data in Figure 19, Figure 20, and Figure 21. According to Figure 19 and Figure 20, both the drug count and interaction count relationship and drug-protein count and interaction count seem to follow a power-law distribution. However, the evidence is not as strong as the previous case (for instance, drug-protein interactions), and a conclusion cannot be drawn in this case. That being said, according to Figure 21, interacting drugs share the same target protein at a rate of 6%. This result is meaningful since drug pairs generally do not have the same target as drug targets are specific to drugs, and their commonality is low. Contrary to this, interacting drugs have the same enzyme at a rate of 64%, have the same carrier at a rate of 70%, have the same transporter at 56%.

According to Figure 22, drug interaction possibility is positively correlated with the drug ATC level. That makes sense since the higher the ATC level, the higher the likelihood of drug interactions, as the similarity between drugs also increases.

Another DDI classification, as shown in Table 8 in the previous section, is viable to the description information of interacting drug pairs. Based on the mentioned classification, we conclude that interacted drug pairs in the first group having interaction description “The risk or severity of adverse effects can be increased when Drug A is combined with Drug B” have more adverse drug-drug interactions (ADDIs) potential compared to interactions in the other four groups.

5.5 Conclusion

The primary purpose of this study is to emphasize the effect of genetic variability on DDI occurrence, which we conclude is one of the most potent factors on drug response variability. We designed a relational database that stores DDI and genetic variability associations.

The designed database was implemented on top of the PostgreSQL relational database management system (RDBMS). The data content has DDI, DGI, and DDGI data extracted from different knowledge bases, including DrugBank, PharmGKB, Ensembl, KEGG Drug, and ONC. The content of this study was shared via a user-friendly web interface. The web interface, named DDGICat Browser, was developed with the R Shiny package. DDGICat Browser enables users to search for a particular drug, gene, SNP, interacting drug pairs, and drug-gene interaction data. Additionally, It enables viewing interacting drug pairs which share the same genetic materials.

The designed database (DDGICat) was tested on a case study on a sample disease (Acute Coronary Syndrome (ACS)) via the developed visual interface (DDGICat Browser). Drug pairs having interaction possibilities in associated chromosomes, genes, proteins, and SNP details were shared within the case study.

In addition, statistical analysis results obtained from the developed database (DDGICat) were shared. Among these results, it has been shown that the drug-associated protein (target, enzyme, carrier, transporter) distribution per drug follows a power-law distribution, meaning that drugs and associated proteins are a scale-free network. Based on the obtained results, it is possible to say that there are “hub” drugs and drug proteins.

5.6 Future Studies

The possible additions and improvements to this study are summarized below:

- More metadata relevant to data quality and provenance should be associated.
- More data containing clinically oriented interaction severity information can be imported into the existing DDI records.
- A machine learning model may be generated based on the collected data to predict the unknown DDI, DGI, and DDGI records.

REFERENCES

- [1] N. Shehab, M. C. Lovegrove, A. I. Geller, K. O. Rose, N. J. Weidle, and D. S. Budnitz, “Us emergency department visits for outpatient adverse drug events, 2013-2014,” *Jama*, vol. 316, no. 20, pp. 2115–2125, 2016.
- [2] M. L. Becker, M. Kallewaard, P. W. Caspers, L. E. Visser, H. G. Leufkens, and B. H. Stricker, “Hospitalisations and emergency department visits due to drug–drug interactions: a literature review,” *Pharmacoepidemiology and drug safety*, vol. 16, no. 6, pp. 641–651, 2007.
- [3] S. L. Van Driest, Y. Shi, E. A. Bowton, J. S. Schildcrout, J. F. Peterson, J. Pulley, J. C. Denny, and D. M. Roden, “Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing,” *Clinical Pharmacology & Therapeutics*, vol. 95, no. 4, pp. 423–431, 2014.
- [4] M. R. Knisely, J. S. Carpenter, C. B. Draucker, T. Skaar, M. E. Broome, A. M. Holmes, and D. Von Ah, “Cyp2d6 drug-gene and drug-drug-gene interactions among patients prescribed pharmacogenetically actionable opioids,” *Applied Nursing Research*, vol. 38, pp. 107–110, 2017.
- [5] R. K. Thirumaran, J. W. Heck, and B. T. Hocum, “Cyp450 genotyping and cumulative drug–gene interactions: an update for precision medicine,” 2016.
- [6] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. P. Tatonetti, S. Vi-lar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, *et al.*, “Toward a complete dataset of drug–drug interaction information from publicly available sources,” *Journal of biomedical informatics*, vol. 55, pp. 206–217, 2015.
- [7] W. H. Organization *et al.*, *Principles for pre-clinical testing of drug safety: re-port of a WHO Scientific Group [meeting held in Geneva from 21 to 26 March 1966]*. World Health Organization, 1966.
- [8] E. Ravina, *The evolution of drug discovery: from traditional medicines to mod-ern drugs*. John Wiley & Sons, 2011.
- [9] W. Bynum, *The history of medicine: a very short introduction*, vol. 191. Oxford University Press, 2008.
- [10] K. Tripathi, *Essentials of medical pharmacology*. JP Medical Ltd, 2013.

- [11] S. Scheindlin, “Our man in dorpat: Rudolf buchheim and the birth of pharmacology,” *Molecular interventions*, vol. 10, no. 6, p. 331, 2010.
- [12] M. Rowland, T. Tozer, H. Derendorf, and G. Hochhaus, “Clinical pharmacokinetics and pharmacodynamics. 2011,” 2011.
- [13] H. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan, “Advancing drug discovery via artificial intelligence,” *Trends in pharmacological sciences*, vol. 40, no. 8, pp. 592–604, 2019.
- [14] R. Seifert, *Basic knowledge of pharmacology*. Springer, 2019.
- [15] Q. Ma and A. Y. Lu, “Pharmacogenetics, pharmacogenomics, and individualized medicine,” *Pharmacological reviews*, vol. 63, no. 2, pp. 437–459, 2011.
- [16] K. J. Karczewski, R. Daneshjou, and R. B. Altman, “Chapter 7: pharmacogenomics,” *PLoS computational biology*, vol. 8, no. 12, p. e1002817, 2012.
- [17] S. M. M. Alsanosi, C. Skiffington, and S. Padmanabhan, “Pharmacokinetic pharmacogenomics,” 2014.
- [18] L. McCallum, S. Lip, and S. Padmanabhan, “Pharmacodynamic pharmacogenomics,” 2014.
- [19] T. N. Tozer and M. Rowland, *Essentials of pharmacokinetics and pharmacodynamics*. 2016.
- [20] J. Doan, H. Zakrzewski-Jakubiak, J. Roy, J. Turgeon, and C. Tannenbaum, “Prevalence and risk of potential cytochrome p450-mediated drug-drug interactions in older hospitalized patients with polypharmacy,” *Annals of Pharmacotherapy*, vol. 47, no. 3, pp. 324–332, 2013.
- [21] R. L. Maher, J. Hanlon, and E. R. Hajjar, “Clinical consequences of polypharmacy in elderly,” *Expert opinion on drug safety*, vol. 13, no. 1, pp. 57–65, 2014.
- [22] M. C. S. Rodrigues and C. d. Oliveira, “Drug-drug interactions and adverse drug reactions in polypharmacy among older adults: an integrative review1,” *Revista latino-americana de enfermagem*, vol. 24, 2016.
- [23] L. Magro, U. Moretti, and R. Leone, “Epidemiology and characteristics of adverse drug reactions caused by drug–drug interactions,” *Expert opinion on drug safety*, vol. 11, no. 1, pp. 83–94, 2012.
- [24] G. L. Kedderis, “Pharmacokinetics of drug interactions,” *Advances in Pharmacology*, vol. 43, pp. 189–203, 1997.
- [25] S.-M. Huang, R. Temple, D. Throckmorton, and L. Lesko, “Drug interaction studies: study design, data analysis, and implications for dosing and labeling,” *Clinical Pharmacology & Therapeutics*, vol. 81, no. 2, pp. 298–304, 2007.

- [26] L. S. Elliott, J. C. Henderson, M. B. Neradilek, N. A. Moyer, K. C. Ashcraft, and R. K. Thirumaran, “Clinical impact of pharmacogenetic profiling with a clinical decision support tool in polypharmacy home health patients: A prospective pilot randomized controlled trial,” *PloS one*, vol. 12, no. 2, p. e0170905, 2017.
- [27] G. Koren, J. Cairns, D. Chitayat, A. Gaedigk, and S. J. Leeder, “Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother,” *The Lancet*, vol. 368, no. 9536, p. 704, 2006.
- [28] A. G. Motulsky, “DRUG REACTIONS, ENZYMES, AND BIOCHEMICAL GENETICS,” pp. 5–7, 1957.
- [29] K. K. Jain and K. Jain, *Textbook of personalized medicine*. Springer, 2009.
- [30] A. Marshall, “Genset–abbott deal heralds pharmacogenomics era,” *Nature biotechnology*, vol. 15, no. 9, pp. 829–830, 1997.
- [31] P. Verbeurgt, T. Mamiya, and J. Oesterheld, “How common are drug and gene interactions? prevalence in a sample of 1143 patients with cyp2c9, cyp2c19 and cyp2d6 genotyping,” *Pharmacogenomics*, vol. 15, no. 5, pp. 655–665, 2014.
- [32] M. Ingelman-Sundberg, “The human genome project and novel aspects of cytochrome p450 research,” *Toxicology and applied pharmacology*, vol. 207, no. 2, pp. 52–56, 2005.
- [33] K. Raja, M. Patrick, J. T. Elder, and L. C. Tsoi, “Machine learning workflow to enhance predictions of adverse drug reactions (adrs) through drug-gene interactions: application to drugs for cutaneous diseases,” *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.
- [34] B. Percha, Y. Garten, and R. B. Altman, “Discovery and explanation of drug-drug interactions via text mining,” in *Biocomputing 2012*, pp. 410–421, World Scientific, 2012.
- [35] S. Dere and S. Ayvaz, “Prediction of drug–drug interactions by using profile fingerprint vectors and protein similarities,” *Healthcare informatics research*, vol. 26, no. 1, pp. 42–49, 2020.
- [36] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, and X. Li, “Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data,” *BMC bioinformatics*, vol. 18, no. 1, pp. 1–12, 2017.
- [37] S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R. Rabadan, and C. Friedman, “Drug—drug interaction through molecular structure similarity analysis,” *Journal of the American Medical Informatics Association*, vol. 19, no. 6, pp. 1066–1074, 2012.

- [38] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Label propagation prediction of drug-drug interactions based on clinical side effects," *Scientific reports*, vol. 5, no. 1, pp. 1–10, 2015.
- [39] R. Ferdousi, R. Safdari, and Y. Omid, "Computational prediction of drug-drug interactions based on drugs functional similarities," *Journal of biomedical informatics*, vol. 70, pp. 54–64, 2017.
- [40] R. Safdari, R. Ferdousi, K. Aziziheris, S. R. Niakan-Kalhari, and Y. Omid, "Computerized techniques pave the way for drug-drug interaction prediction and interpretation," *BioImpacts: BI*, vol. 6, no. 2, p. 71, 2016.
- [41] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.
- [42] F. Storelli, C. Samer, J.-L. Reny, J. Desmeules, and Y. Daali, "Complex drug–drug–gene–disease interactions involving cytochromes p450: systematic review of published case reports and clinical perspectives," *Clinical pharmacokinetics*, vol. 57, no. 10, pp. 1267–1293, 2018.
- [43] D. Brixner, E. Biltaji, A. Bress, S. Unni, X. Ye, T. Mamiya, K. Ashcraft, and J. Biskupiak, "The effect of pharmacogenetic profiling with a clinical decision support tool on healthcare resource utilization and estimated costs in the elderly exposed to polypharmacy," *Journal of medical economics*, vol. 19, no. 3, pp. 213–228, 2016.
- [44] H. Zakrzewski-Jakubiak, J. Doan, P. Lamoureux, D. Singh, J. Turgeon, and C. Tannenbaum, "Detection and prevention of drug–drug interactions in the hospitalized elderly: utility of new cytochrome p450–based software," *The American journal of geriatric pharmacotherapy*, vol. 9, no. 6, pp. 461–470, 2011.
- [45] M. A. Malki and E. R. Pearson, "Drug–drug–gene interactions and adverse drug reactions," *The pharmacogenomics journal*, vol. 20, no. 3, pp. 355–366, 2020.
- [46] M. Tod, C. Nkoud-Mongo, and F. Gueyffier, "Impact of genetic polymorphism on drug-drug interactions mediated by cytochromes: a general approach," *The AAPS journal*, vol. 15, no. 4, pp. 1242–1252, 2013.
- [47] M. A. Bahar, D. Setiawan, E. Hak, and B. Wilffert, "Pharmacogenetics of drug–drug interaction and drug–drug–gene interaction: a systematic review on cyp2c9, cyp2c19 and cyp2d6," *Pharmacogenomics*, vol. 18, no. 7, pp. 701–739, 2017.
- [48] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, "Drugbank 5.0: a major update

- to the drugbank database for 2018,” *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.
- [49] M. Whirl-Carrillo, E. M. McDonagh, J. Hebert, L. Gong, K. Sangkuhl, C. Thorn, R. B. Altman, and T. E. Klein, “Pharmacogenomics knowledge for personalized medicine,” *Clinical Pharmacology & Therapeutics*, vol. 92, no. 4, pp. 414–417, 2012.
 - [50] C. F. Thorn, T. E. Klein, and R. B. Altman, “Pharmgkb: the pharmacogenomics knowledge base,” in *Pharmacogenomics*, pp. 311–320, Springer, 2013.
 - [51] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, “Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart,” *Nature protocols*, vol. 4, no. 8, pp. 1184–1191, 2009.
 - [52] M. Kanehisa and S. Goto, “Kegg: kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.
 - [53] S. Phansalkar, A. A. Desai, D. Bell, E. Yoshida, J. Doole, M. Czochanski, B. Middleton, and D. W. Bates, “High-priority druggedrug interactions for use in electronic health records,” *Journal of the American Medical Informatics Association*, vol. 19, pp. 735–743, 2012.
 - [54] L.-H. Huang, Q.-S. He, K. Liu, J. Cheng, M.-D. Zhong, L.-S. Chen, L.-X. Yao, and Z.-L. Ji, “Adrecs-target: target profiles for aiding drug safety research and application,” *Nucleic acids research*, vol. 46, no. D1, pp. D911–D917, 2018.
 - [55] K. C. Cotto, A. H. Wagner, Y.-Y. Feng, S. Kiwala, A. C. Coffman, G. Spies, A. Wollam, N. C. Spies, O. L. Griffith, and M. Griffith, “Dgidb 3.0: a redesign and expansion of the drug–gene interaction database,” *Nucleic acids research*, vol. 46, no. D1, pp. D1068–D1073, 2018.
 - [56] Y. Yu, Y. Wang, Z. Xia, X. Zhang, K. Jin, J. Yang, L. Ren, Z. Zhou, D. Yu, T. Qing, *et al.*, “Premedkb: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs,” *Nucleic acids research*, vol. 47, no. D1, pp. D1090–D1101, 2019.
 - [57] K. Lee, B. Kim, Y. Choi, S. Kim, W. Shin, S. Lee, S. Park, S. Kim, A. C. Tan, and J. Kang, “Deep learning of mutation-gene-drug relations from the literature,” *BMC bioinformatics*, vol. 19, no. 1, pp. 1–13, 2018.
 - [58] F. Moreau, N. Simon, M. Tod, B. Décaudin, and P. Odou, “Ddi-predictor: A novel clinical pharmacy decision-making tool for dose adaptation?,” 2019.
 - [59] S. Qian, S. Liang, and H. Yu, “Leveraging genetic interactions for adverse drug-drug interaction prediction,” *PLoS computational biology*, vol. 15, no. 5, p. e1007068, 2019.

- [60] M. Ali and A. Ezzat, *DrugBank Database XML Parser*. Dainanahan, 2020. R package version 1.2.0.
- [61] B. Momjian, “Postgresql: introduction and concepts addison-wesley,” *New York*, 2001.
- [62] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, *shiny: Web Application Framework for R*, 2021. R package version 1.6.0.



APPENDIX A

PROJECT CODES

All the R and SQL code created for this thesis is available at:
<https://github.com/aycomp/DDGICat>