BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

DOCTORAL THESIS

Brno, 2021 Mgr. Pavlína Cicková

BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING
ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

METHODS FOR PREDICTING DRUG SIDE EFFECTS IN SILICO
METODY IN SILICO PREDIKCE NEŽÁDOUCÍCH ÚČINKŮ LÉČIV

DOCTORAL THESIS
DIZERTAČNÍ PRÁCE

AUTHOR — **Mgr. Pavlína Cicková**
AUTOR PRÁCE

SUPERVISOR — **prof. Ing. Ivo Provazník, Ph.D.**
ŠKOLITEL

BRNO 2021

ABSTRACT

Drug discovery is a field of contemporary science, which has encompassed the use of various computational methods. Wet lab approaches are costly and time-consuming and hence, *in silico* methods play an important role. Notwithstanding the progress of computational techniques applied in drug discovery in the last few decades, the great majority of the investigational compounds still do not succeed in reaching the final approval stage. Not only for this reason state-of-the-art drug design strategies focus on reinvestigating already approved drugs and drug similarity analyzes are crucial to consider. This work presents the development and application of a set of workflows created within the KNIME Analytics Platform which implements an approach using machine-learning methods for drug side effect prediction. The presented set of workflows deals with data retrieval, pre-processing, similarity metrics computation and data exploratory analysis. Consequently, classification models are applied to predict specific side effects of drugs. The prediction is based on similarity-based techniques. Structural and other similarities of approved drug molecules were used to train the decision tree models for the prediction of potential drug-side effect associations. The main advantage of the work is the re-usability of the applied techniques. Our set of workflows provides an environment allowing for new research questions in terms of drug similarity to be addressed. Moreover, as the workflows created within KNIME Analytics Platform provide a user-friendly graphical interface, users do not require any advanced experience in machine learning or programming to perform their studies using the designed workflows.

KEYWORDS

bioinformatics; big data; data integration; data mining; data processing; data science; drug discovery; drug design; drug interactions; chemoinformatics; *in silico* prediction; KNIME; machine learning; similarity; side effects; workflow

ABSTRAKT

Vývoj a výzkum léčiv je oblastí současné vědy, jejíž nedílnou součástí je i využití výpočetních metod. Z důvodu nákladnosti a časové náročnosti laboratorních přístupů, metody *in silico* sehrávají svou významnou roli. I přes rychlý vývoj výpočetních technik využívaných při vývoji léků, však není drtivá většina zkoumaných molekul v procesu vývoje úspěšná a do schvalovací fáze nepostoupí. Nejen proto se nejmodernější strategie návrhu potenciálních nových léčiv zaměřují na opětovné zkoumání již schválených léků a berou do úvahy i analýzu podobností. Tato práce popisuje vývoj a aplikaci souboru několika workflow, jež byl vytvořen v rámci analytické platformy KNIME a jež implementuje metody strojového učení za účelem predikce nežádoucích účinků léčiv. Součástí prezentovaných workflow je získání dat, jejich předzpracování, výpočet metrik podobností a provedení explorační analýzy. Následně je využito klasifikačních modelů k predikci specifických nežádoucích účinků léčiv. Tato predikce vychází z principů technik založených na podobnosti. K natrénování modelů rozhodovacích stromů pro predikci potenciální asociace nežádoucích účinků s léčivy byly využity strukturní a jiné podobnosti schválených molekul léčiv. Hlavní přínos práce spočívá především v přenositelnosti použitých metod. Soubor workflow je určen k využití jako vhodný nástroj k řešení výzkumných otázek ohledně podobnosti léčiv a jelikož analytická platforma KNIME poskytuje uživatelsky přívětivé grafické rozhraní, není nutné, aby měli uživatelé pokročilé zkušenosti v oblasti strojového učení nebo programování, aby mohli soubor navržených workflow v rámci této platformy pro své analýzy využít.

KLÍČOVÁ SLOVA

bioinformatika; big data; integrace dat; dolování v datech; zpracování dat; datová věda; objev léčiv; návrh léčiv; lékové interakce; chemoinformatika; predikce *in silico*; KNIME; strojové učení; podobnost; nežádoucí účinky; workflow

ROZŠÍŘENÝ ABSTRAKT

Tato práce se zabývá vývojem nástroje pro predikci nežádoucích účinků léčiv. Cílem této práce bylo vytvořit nástroj využívající strojové učení a zaměřit se na využití metod založených na podobnosti molekul léčiv. Na základě získaných znalostí byla zkoumána vhodnost využití některých metrik podobností k predikci nežádoucích účinků.

Práce je členěna do pěti kapitol. V první kapitole je uvedeno téma a motivace práce, jsou formulovány cíle a rovněž je stručně popsáno členění celého textu.

V druhé kapitole jsou představeny teoretické poznatky, které se k tématu práce vztahují. Kapitola nejprve čtenáře seznamuje s problematikou moderního vývoje léčiv a krátce se věnuje charakteristice jeho jednotlivých fází. Následně jsou popsány zdroje dat, které jsou při vývoji léčiv využívány, především je však text kapitoly zaměřen na konkrétní databáze, z nichž byla získána data pro tuto práci. Nedílnou součástí kapitoly je rovněž objasnění konceptu možného výpočtu podobnosti, využití technik strojového učení při vývoji léčiv a také jsou zmíněny studie, které s tématem práce souvisí. V neposlední řadě se kapitola věnuje představení a využití softwaru „KNIME Analytics Platform“, ve kterém byl nástroj vytvořen.

Třetí kapitola popisuje metody použité v práci. Jsou zde formulovány konkrétní kroky, jak byla získána a zpracována data, jak byla stanovena podobnost mezi molekulami léčiv a jak byly sestaveny datasety využité v analýzách. Také je zde popsáno, které nežádoucí účinky byly zkoumány, a jak probíhala explorační analýza dat. Poté jsou zmíněny podrobnosti sestavených modelů rozhodovacích stromů a rovněž jsou popsány metriky vyhodnocení přesnosti predikcí.

Ve čtvrté kapitole jsou představena konkrétní navržená workflow. Tato část práce rovněž diskutuje výsledky analýzy různých datových souborů. Dosažené výsledky jsou vyhodnoceny a vizualizovány pomocí příslušných grafů.

V páté závěrečné kapitole je zhodnoceno dosažení stanovených cílů, jsou zde shrnuty návrhy na další možná rozšíření navržených workflow a také jsou uvedeny limitace nástroje.

Následuje seznam použité literatury, seznam zkratek a seznam příloh. Do příloh je zahrnut souhrn použitých skriptů, datasetů a souborů workflow, a výsledky korelační analýzy. Na konci práce se nachází životopis autorky včetně seznamu publikací, k jejichž vzniku se během svého doktorského studia přičinila.

Výstupem této práce je komplexní soubor workflow pro dolování v datech léčiv a vizualizaci získaných výsledků. K vývoji nástroje bylo využito prostředí globální open-source platformy „KNIME Analytics Platform“. Ta je určena pro různorodé úlohy v oblasti automatického zpracování rozsáhlých dat. Nástroj je velmi efektivní a je hojně využíván pro organizaci a analýzu dat, strojové učení či datovou vědu.

Velkou výhodou této platformy je především její přívětivé uživatelské rozhraní a aktivní vývojářská komunita.

Pro účely této práce bylo vytvořeno několik workflow, která na sebe navazují a tvoří finální soubor 'DISSERTATION_PROJECT.knar'. Lze je využít jako jeden celek či jako samostatná workflow pro vyřešení konkrétní úlohy příslušného kroku datové analýzy. Do workflow jsou implementovány vlastní skripty naprogramované v jazyku R.

Součástí prezentovaných workflow je získání a předzpracování dat z volně přístupných databází DrugBank a SIDER. K výpočtu metrik podobností mezi léčivy byl využit Tanimoto/Jaccard koeficient. Pro účely analýzy bylo sestaveno několik různých datových souborů pomocí navržené filtrovací funkce. Nedílnou součástí je pak explorační analýza zkoumaných dat. K predikci specifických nežádoucích účinků léčiv je využito klasifikačních modelů. K sestavení těchto modelů byla zvolena metoda rozhodovacích stromů. K odhadu přesnosti modelů bylo využito 10násobné křížové validace. Výsledky přesnosti predikce modelů jsou porovnány a vizualizovány pomocí krabicových diagramů.

Dle výsledků byly stanoveny parametry vhodné k predikci nežádoucích účinků na základě podobnosti léčiv. Z výsledků vyplývá, že největší přínos ze studovaných metrik má pro predikci nežádoucích účinků strukturní podobnost molekul. Zvýšené přesnosti modelů bylo dosaženo díky implementaci navržené filtrovací funkce.

Hlavní přínos práce spočívá v přenositelnosti použitých metod. Díky přívětivému uživatelskému rozhraní analytické platformy KNIME a intuitivnímu nastavení jednotlivých uzlů nemusí mít uživatelé pokročilé zkušenosti v oblasti strojového učení nebo programování, aby mohli soubor navržených workflow pro své analýzy využít. Nástroj je určen pro analýzy při vývoji molekul potenciálních léčiv a jeho záměrem je umožnit uživatelům dolováním v datech získat potenciální nové znalosti o léčivech. V kombinaci s jinými přístupy tak nástroj může přispět k zefektivnění vývoje léčiv.

CICKOVÁ, Pavlína. *Methods for predicting drug side effects in silico*. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering, 2021, 130 p. Doctoral thesis. Advised by prof. Ing. Ivo Provazník, Ph.D.

Author's Declaration

Author:	Mgr. Pavlína Cicková
Author's ID:	175408
Paper type:	Doctoral thesis
Academic year:	2021/22
Topic:	Methods for predicting drug side effects *in silico*

I declare that I have written this paper independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the paper and listed in the comprehensive bibliography at the end of the paper.

As the author, I furthermore declare that, with respect to the creation of this paper, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll. of the Czech Republic, Section 2, Head VI, Part 4.

Brno ....

author's signature*

*The author signs only in the printed version.

ACKNOWLEDGEMENT

I am grateful that I could enhance my research skills and participate in many various interdisciplinary projects at the Department of Biomedical Engineering. I would like to thank the supervisor of my thesis Prof. Ing. Ivo Provazník, Ph.D. His motivational guidance was what helped me stay productive and enthusiastic about my work throughout the years of my studies. I would also like to thank my colleagues who have helped create an intellectually stimulating environment and who have become my companions. I am also thankful to my family and friends for support. Especially, I would like to acknowledge my husband Matúš, who encouraged me in my work from the very beginning to the very end.

Contents

List of Figures

List of Tables

1 Introduction

The following chapter presents the main topic of the thesis. The first subchapter focuses on the topic introduction, the second one explains the motivation and the third one presents aims and objectives of the thesis. The last subchapter provides a guide to the chapters of the thesis.

1.1 Topic introduction

Developing efficient medicinal drugs is an enormously expensive and time-consuming process. Nowadays, total costs of a bringing a new drug to market attacks $1-3 billion depending on a therapeutic area (Wouters et al., 2020) and the development takes approximately 10–15 years (Matthews et al., 2016). Moreover, a lot of effort to create novel therapies dies in development. The great majority of the investigational compounds entering clinical trials do not work as expected and hence do not succeed in reaching the final approval stages (C. H. Wong et al., 2019; Dowden et al., 2019). For various reasons, many suggested molecules fail in the later phases of drug discovery after being considered successful in animal tests. The reasons for these failures involve toxicity, undesirable (and unexpected) side effects, a lack of efficacy or a failure to demonstrate value compared to an existing therapy. A reason to fail is a lack of knowledge which can be predicted about the compounds before testing.

In order to increase the success rates of drug development efforts, it is of importance to predict the drug side effects beforehand. Failures in phases II and III of clinical trials are extremely costly (Plenge, 2016) and therefore there are many attempts at trying to decrease the high probability of failure. The issue regarding identifying potential side effects in the early stages and better selecting of candidate molecules for further analysis can be overcome by the novel concepts.

A plethora of computational methods have been developed so far for the purposes of drug discovery. The main reason why computational technologies play such a crucial role in drug development, is their ability to solve the critical problem of which molecule will most likely succeed, and which will most likely fail before they are sent to costly wet-lab testing. Hence, they make the discovery process more time and cost-effective.

Without doubt 'Big Data era' has impacted pharmaceutical drug discovery and development and nowadays, interdisciplinary cooperation is essential. As informatics education is not integrated into many chemistry curricula, workflow based approaches are developed in order to help medical chemists process large amounts of data.

1.2 Motivation and hypothesis

The main goal of this thesis is to develop a simple research tool which can be used in computational drug design attempts for predicting side effects, because improved *in silico* predictions allow for eliminating a number of *in vivo* and *in vitro* experiments. As we hope that exploring the universe of drug similarities can provide valuable information, the purpose of the designed tool would be to predict drug side effects based on drug similarities and specify those drugs which are worthy of further study in terms of their association to specified side effects. Our hypothesis states that drug molecules that share more similarities will demonstrate higher number of shared side effects compared to those drug molecules with fewer similarities between each other.

This work seeks to design a reliable, comprehensive, easy-to-understand, and easy-to-use framework for predicting side effects. The purpose of such framework is to use a novel combination of available data and provide a novel solution to be used in the drug discovery process. The created pipeline would combine data retrieval and processing, exploration, analysis, visualization, and reporting the results. The framework will be able to process up-to-date information from freely available web services. The proposed customizable workflow could be further used in computational drug-development attempts by bioinformatics researchers and it would be available for modifications. Advantages of such a framework include saving time, standardization, and research reproducibility.

1.3 Aims and objectives

The main aims and objectives of this thesis cover the following:

1. **To develop a pipeline for semi-automatic side effect predictions.**

    - A customizable *in silico* drug discovery pipeline with the potential for accelerating preclinical stages will be developed. The set of workflows will be created in free data management software.
    - Data deployment, processing and exploration steps will be included, as well as inspecting the data by a similarity-based approach. A machine learning algorithm and statistical methods will be implemented in order to guide predicting drug side effects attempts.
    - The developed tool will be simple to use for non-experts. It will assist them in analyzing of the large amounts of data and determining drug side effects.

2. **To apply the developed set of workflows to real data.**

   - The set of designed workflows will be applied for predicting side effect using integrated data from open-source databases.
   - Various drug data information will be collected, filtered, and explored, and consequently used for drug similarity metrics calculation.
   - Calculated metrics will be used as features to train machine learning models. The performance of the models will be scored and the features will be observed in terms of enhancing the performance of the models for predicting side effects.
   - Predicting side effects via machine learning models will be classified using different evaluation metrics.

1.4 Guide to the chapters

The intent of this subchapter is to briefly describe the thesis structure. This thesis is structured in the following order:

The first chapter includes a brief introduction to the topic of the thesis, its motivation and proposed hypothesis. The aims and objectives are discussed in this chapter as well.

It is the purpose of the second chapter to provide an introduction to the theoretical background. In this section, we explain the basic terminologies used in this work. Moreover, we mention the types of used databases as well as the techniques applied in the work.

Third chapter covers the research methodologies applied in this dissertation. It is focused on explaining of dataset construction, data preparation, and model development. Furthermore, this chapter explains the applied algorithms and evaluation metrics.

The fourth chapter gives an overview of the results and their discussion. Here, we present the designed workflows in detail and demonstrate the experimental results which validate the effectiveness of our approach.

In the fifth chapter of the thesis, we draw conclusions and provide ideas for future research attempts. The limitations for consideration and possible further improvements are put together here as well.

At the end of the thesis, the reader can find a biography, symbols and abbreviations, and a list of appendices which provides details on additional files including author's vita and a list of publications.

2 Theoretical background

The drug development process has changed dramatically over the past century. The following chapter briefly introduces the context of modern drug development. Next, selected drug databases are presented. This is followed by an introduction to similarity-based and machine-learning approaches in drug discovery and related research reviews. In particular, some computational approaches which have been applied for identifying drug side effects are mentioned. Finally, the workflow management software applied in our work is presented.

2.1 Drug discovery and development in the era of Big Data

In the past, drug discovery was based on isolating molecules from natural sources or synthesizing new substances and testing them for treatment. To date drug discovery has changed in the face of rapid advances in technology, as well as new approaches developed by scientists. It has significantly improved with the knowledge that diseases are controlled at the molecular and physiological level, by understanding the shape of molecules at the atomic level, or with information about the human genome (Chast, 2008). The size of chemical space of drug-like compounds has been estimated as 10^{33} (Polishchuk et al., 2013), therefore, as one can imagine, the process of drug discovery is much like looking for a needle in a haystack.

The developent of a medicinal agent is a lengthy inter-disciplinary process which combines aspects of bioinformatics, chemoinformatics, structural biology, or structure-based drug design. Each of these aspects plays a key role in various stages. The steps involved in the traditional drug discovery and development process are summarized in the picture below (Fig. 2.1).

The process begins with target identification and validation. In general, a medical drug is regarded as a small organic compound which interacts with its target (DNA, proteins, enzymes, or pathways) in the human body and boosts or inhibits its function which is important for the disease progression. Therefore, the initial fundamental step in the drug development process is to identify the biological origin of a disease and define and validate the right drug target or a combination of targets. There are many different approaches for hunting targets in drug discovery. These can include phenotypic screening, imaging, biomarkers, gene association studies, chemo proteomics, experiments on transgenetic organisms, and many other methods (Schenone et al., 2013; George et al., 2017).

Once the target is proposed and validated, it is followed by hits identification.

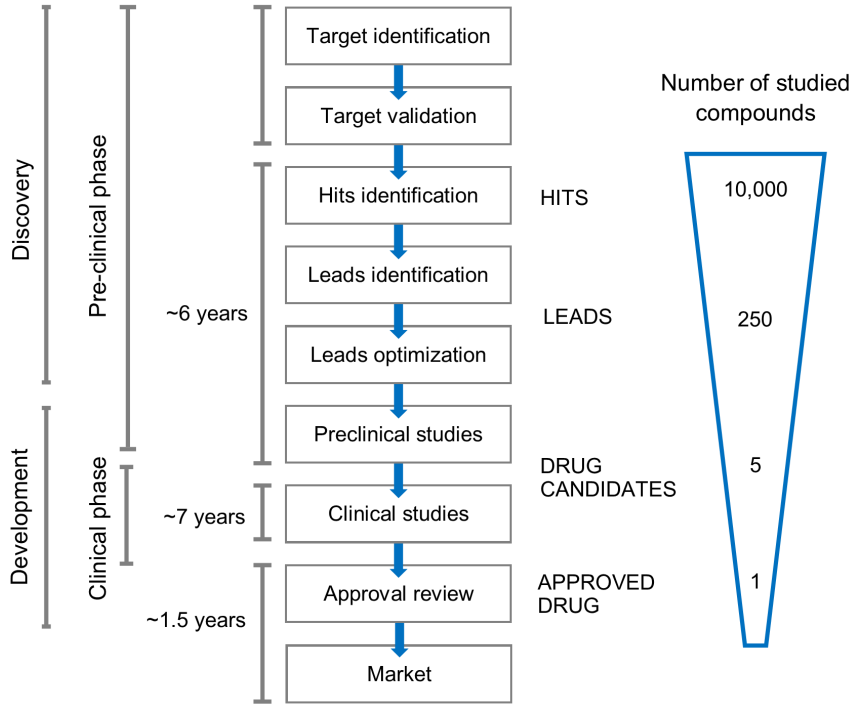


Fig. 2.1: The steps and the success rate of traditional drug development. Early drug discovery includes target identification, target validation, hit discovery, lead identification and lead optimization. This is followed by preclinical and clinical studies and the last step before releasing drug in the market is its approval. The whole process takes approximately 10–15 years and it is estimated that only 1 tested compound out of 10,000 makes it to the market. Adapted from Hughes et al., 2011; Matthews et al., 2016; Gao et al., 2010.

The step aims to screen of small organic molecules libraries to identify so-called 'hits'—potential molecules which would selectively interact with the target and stimulate the desired effect. This step can be concluded with several *in vivo* approaches as high-throughput screening, fragments screening, or physiological screening (Keserű et al., 2006). However, it can also be conducted computationally via a variety of *in silico* virtual screening methods. These computer simulations provide a deeper insight into the complex functioning (Goodnow, 2006).

The lead identification and optimization step is represented by screening small molecules which aims to determine candidates satisfying specific drug properties (M. Wong et al., 2017). The step includes predicting the potential side effects of the drugs or their metabolites. Valuable data are obtained during safety pharmacology, genetic toxicology, chronic toxicology, ADME/PK (absorption, distribution, metabolism, excretion, and pharmacokinetics), and further studies which are applied to investigating the potential undesirable effects of the tested molecules.

Next, drug candidate molecules need to be tested in expensive preclinical and clinical trials to be evaluated as safe and effective before their approval (Umscheid et al., 2011). Preclinical pharmacology studies are undertaken *in vitro* (e.g., cell) and *in vivo* (in suitable animal models). In addition, studies are performed *in silico* (via computer models). Then follow human clinical trials which consist of three main phases. The first 'in-human' testing is carried out during phase I, when studies are conducted with a small number of healthy human volunteers (up to a few dozen tested subjects). Phase II consists of testing larger numbers of patients (up to a few hundred tested subjects). Finally, phase III covers comparative trials during which a large number of patients (up to a few thousand tested subjects), multiple countries (populations), and comparisons with current treatment are included. The phases of clinical studies are summarized in a table below (Tab. 2.1). After passing the studies successfully, the drug can be approved by authorities and launched in the market.

The drug approval process is different for each country. In the United States, the largest market globally for pharmaceutical sales, it is the Food and Drug Administration (FDA) agency which approves drugs (`https://www.fda.gov`). In the European Union, European Medicines Agency (EMA) is in charge of this responsibility (`https://www.ema.europa.eu`). The launch of the drug is followed by post-marketing surveillance or post-authorization safety studies (safety monitoring), also called phase IV clinical trials, during which the adverse effects observed on the prescribed population are recorded (Berlin et al., 2008; Tubach et al., 2011). The safety of an approved drug is monitored and reported as long as it is on the market and drugs can be withdrawn for safety reasons whenever.

As mentioned above, computational (*in silico*) techniques have become an integral part of modern drug discovery and to date there are many various computational approaches applied during the process (Katsila et al., 2016; Begam et al., 2012). The applied methods range from ligand-based or receptor-based methods, to gene ontology and literature mining. The use of the techniques is undisputable. They are present in all stages of drug development from the preclinical discovery stage to the late stage of clinical development. Computational methods have great importance, as they can speed up the whole drug development process and decrease the financial requirements in contrast to experimental testing. Reducing (or even replacing) animal testing is one of their most crucial benefits (Knowles et al., 2003). Impact of Big Data in drug discovery attempts has been discussed in a recent publication (Bajorath et al., 2016), in which the scientists point out various challenges need to be overcome.

Although computational drug discovery has a long history (Drews, 2000), many methods have been developed in recent years. They are used to perform similarity search, machine learning, or statistical approaches. With machine learning and

Tab. 2.1: Clinical studies summary. The side effects identification is covered in stages of preclinical and clinical trials, and it continues during post-approval monitoring (Umscheid et al., 2011).

Timing	**Trial phase**	**Usual number of tested subjects**	**Studied data**
Before approval	I	20–100 healthy human subjects	Safety in humans, maximum tolerated dose, pharmacokinetics, pharmacodynamics, drug-drug interactions
Before approval	II	100–300 patients with the disease, condition	Efficacy at treating diseases, more information on safety, different dosing, control arm, comparing to a standard therapy
Before approval	III	300–3,000 patients	Larger scale studies, diverse population, more information on the efficacy and safety, comparing to a standard therapy or a placebo, randomization, blinding strategies
After approval	IV	Large numbers, diverse populations of patients	Long term safety, less common adverse reactions

data mining approaches, we can investigate what makes a compound a good target or a good drug. In other words, the artificial intelligence algorithms can find an important application in drug development – they can help select drug candidates more reliably.

The technology and tools are continually evolving and there is a wide range of software and a plethora of drug-related data available which is used for *in silico* drug design today. A comprehensive list of open-source molecular modeling tools can be found on the following link: `https://opensourcemolecularmodeling.github.io`

2.2 Drug and side effect databases

There are a plethora of freely accessible databases that provide useful information on drug/target compounds. In the table below, you will find a list of selected databases with a short description of provided data (Tab. 2.2). In this work we use data available from DrugBank (Wishart, 2006) and SIDER databases (Kuhn et al., 2016). These databases have been selected as comprehensive data sources commonly used for scientific research purposes.

The DrugBank database (available at `https://go.drugbank.com`) is a valuable and well-established, freely available resource which contains heterogeneous data on drugs, including known molecular targets, activity in humans, drug targets se-

Tab. 2.2: Selected databases used in drug discovery

Database name	**Data description**	**Website**
BindingDB (Xi Chen et al., 2001, Liu et al., 2007)	measured binding affinities	`https://www.bindingdb.org`
Dictionary of Natural Products	natural products	`https://dnp.chemnetbase.com`
DrugBank (Wishart, 2006)	drug related data	`https://go.drugbank.com`
ClinicalTrials	clinical studies	`https://www.clinicaltrials.gov`
ChemSpider (*ChemSpider*, [n.d.])	chemical structures	`http://www.chemspider.com`
ChEMBL (Gaulton et al., 2012)	chemical structures and bioactivities	`https://www.ebi.ac.uk/chembl/`
Crystallography Open Database	crystal structures of organic, inorganic, metal-organics compounds and minerals	`https://www.crystallography.net`
MATADOR (Manually Annotated Target and Drug Online Resource) (Gunther et al., 2007)	drug-target interactions	`http://matador.embl.de`
MarinLit	marine natural products	`https://marinlit.rsc.org`
PDB (Protein Data Bank) (Bernstein et al., 1977)	3D structures of biomolecules	`https://www.rcsb.org`
SIDER (Side Effect Resource) (Kuhn et al., 2016)	drugs side effects	`http://sideeffects.embl.de`
SureChemBL (Papadatos et al., 2016)	chemical data extracted from the patent literature	`https://www.surechembl.org`
Therapeutic Target Database (Y. H. Li et al., 2018)	therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs	`http://db.idrblab.net/ttd/`
Traditional Chinese Medicine	small molecules based on traditional Chinese medicine	`http://tcm.cmu.edu.tw`
Zinc (Irwin et al., 2005)	commercially available compounds for virtual screening	`https://zinc.docking.org`

quence, structure, or pathway information. It is updated daily, and its downloads are released quarterly. A total number of 14,460 drugs are in the latest DrugBank release (version 5.1.7, released on July 2, 2020) (`https://go.drugbank.com/stats`) and they are organized into the following categories: approved (drugs approved in North America, Europe and Asia), experimental, biotech (drugs with a biological origin–therapeutic proteins, peptides, vaccines, allergenics, blood components, gene therapies etc.), nutraceutical (nutritional supplements), withdrawn (drugs withdrawn due to safety and toxicity issues), illicit (banned drugs). The number of drugs in each of these categories is shown below (Fig. 2.2).

According to the information at the DrugBank website, the approved drugs in the database represent drugs which have passed the approval process anywhere, at least once at some point in time. The list of approved drugs includes also already withdrawn drugs, even when a drug had to be discontinued since receiving this status. The total number of drugs represents a sum of small molecule drugs and biotech drugs.

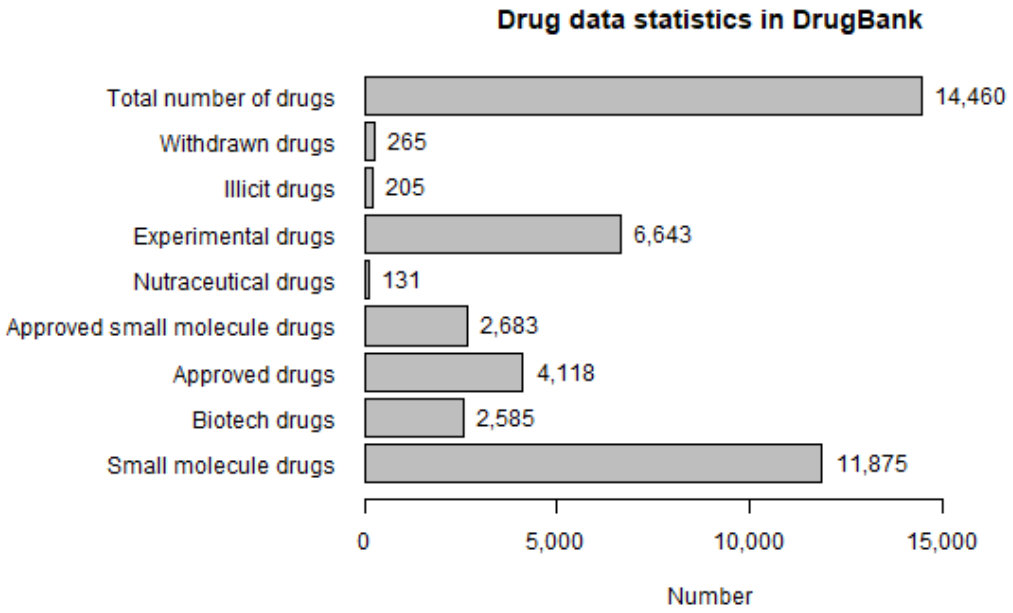


Fig. 2.2: DrugBank statistics (version 5.1.7, released on July 2, 2020)

Yet, drugs in DrugBank database can also be categorized in another way – as small molecule drugs and biotech drugs. Small molecule drugs are molecules with well-defined and a relatively simple structure which are produced by chemical synthesis and have low molecular weight (below 900 daltons). However, within the DrugBank database, even some larger molecules are regarded as small molecule drugs. Biologics are large molecule drugs which have a much more complex structure than small molecule drugs. Biologics are produced by living cells or organisms.

In total there are 11,875 small molecule drugs available in the DrugBank database currently, of which 2,683 of them are approved (Fig. 2.3). The main reason why approved drugs are withdrawn from the market are side effects.

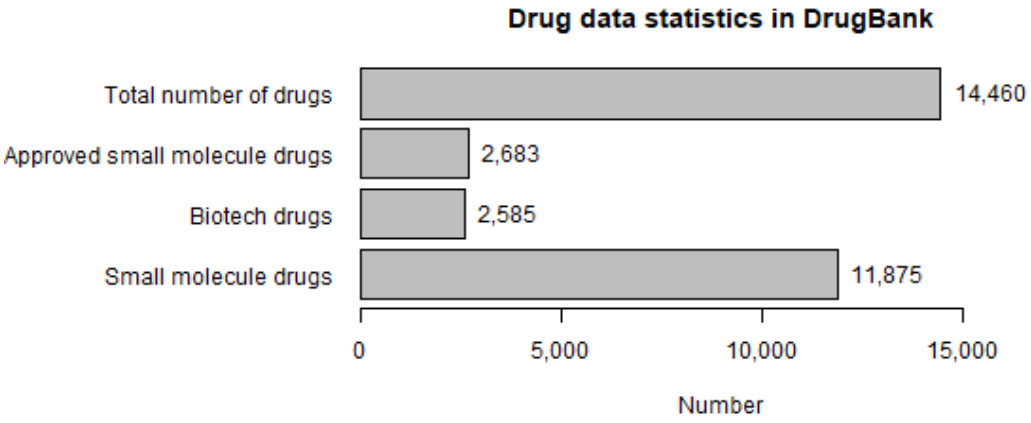


Fig. 2.3: Drug types in DrugBank (version 5.1.7, released on July 2, 2020)

Drug adverse reactions (also called adverse drug events or side effects) is defined as an unexpected, unintended, harmful reaction from a medicine which can endanger patients. It can result in treatment discontinuation, hospitalization, permanent harm, disability, or death.

There are various reasons why adverse drug reactions occur. Many of them are caused by drug-drug interactions as the possibility of adverse reactions increases if multiple drugs are co-administrated inappropriately (Benton et al., 2011). This is mostly the case of patients with complex diseases or several medical conditions who receive multiple therapeutics at the same time. The action of one drug can alter the pharmacological effect of another one. The most frequently co-administered drugs include medications used to treat high blood pressure, heart, psychotropic drugs, or antibiotics. Moreover, adverse drug reactions can also result from different off-target drug reactions when the drug is not aimed at the main target.

As previously-mentioned in section 2.1 regarding the drug discovery and development process, side effect data are not only collected during preclinical and clinical trials, they are also monitored within post marketing surveillance after the drug is approved and released on the market. The Side Effect Resource (SIDER) (available at `http://sideeffects.embl.de`) is a well-known public database which aggregates information on the side effects of marketed drugs (Kuhn et al., 2016). It provides data on adverse reactions which are obtained from public documents and package inserts. The information covers side effect frequency, drug and side effect classifications, or links to further information. In total, the latest release of SIDER database (version 4.1, released on October 21, 2015) contains 139,756 drug–side effect associations corresponding to 5,868 side effects of 1,430 drugs (Fig. 2.4). All

of these drugs are FDA-approved. The database uses internationally established Medical Dictionary for Regulatory Activities (MedDRA) terminology (Brown et al., 1999) to label drugs.

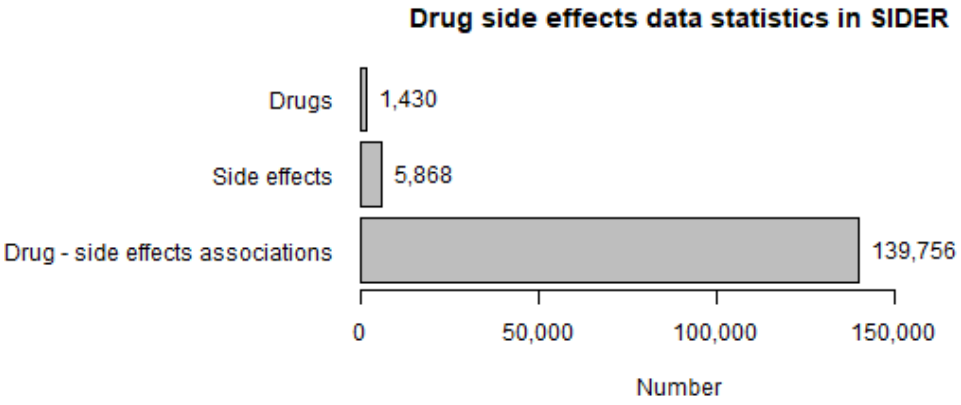


Fig. 2.4: SIDER 4.1 statistics (version 4.1, released on October 21, 2015)

Side effects can be categorized according to Council for International Organization of Medical Sciences (CIOMS) standard recommendations (CIOMS, 1995) available at `https://basicmedicalkey.com/cioms` that suggest side effect frequency classification as given in table below (Tab. 2.3).

Tab. 2.3: CIOMS side effect frequency convention

Side effect classification	**Frequency of patients with the side effect**
Very common	$\geq 1/10$ ($\geq 10\,\%$)
Common (frequent)	$\geq 1/100$ and $< 1/10$ ($\geq 1\%$ and $< 10\,\%$)
Uncommon (infrequent)	$\geq 1/1000$ and $<1/100$ ($\geq 0.1\,\%$ and $< 1\,\%$)
Rare	$\geq 1/10000$ and $< 1/1000$ (≥ 0.01 and $< 0.1\,\%$)
Very rare	$< 1/10000$ ($< 0.01\,\%$)

2.3 Drug similarity data processing

Similarity-based approaches are widely used in many research areas and they have been embedded as a key concept in drug discovery research for a long time (Bender et al., 2004). Molecular similarity can be defined as a 'measure of the degree of overlap between a pair of molecules in some property space' (Allen et al., 2001; Good et al., 2002) and it can be perceived in many different ways, e.g. chemical similarity, molecular similarity, or similarity in biological activity (Maggiora et al.,

2014). The basic assumption is that similar compounds should have similar properties (A.H.-L. et al., 1992).

The concept of molecular similarity has been widely used especially in the early stages of the drug discovery process. It has been suggested that drug molecules with similar chemical structures tend to have similar biological effects (Martin et al., 2002). A number of studies have found that a chemical similarity between two compounds indicates that they could share a target. It has been shown that drug molecules with more common target proteins have a higher degree of similarity (Xing Chen et al., 2016). A study of the interactome network suggested predicting drug-target interactions based on the interactions of the majority of the network neighbors (Z.-C. Li et al., 2016).

The term 'molecular fingerprint' refers to a simplified molecule representation. It is a fixed-length vector comprising of binary digits which correspond to a specific property of the molecule. In the case of structural fingerprints, the properties refer to structural properties of the molecule, such as functional groups, C-chains, ring structures, or the number of bonds and atoms. Each bit in the vector indicates the presence or absence of a specific molecular feature (Fig. 2.5).

Since there is no universal definition of 'molecular similarity', there are a plethora of different fingerprints which can represent relevant structural features in various ways (Morgan fingerprint, Atom-Pair fingerprint, Topological-Torsion fingerprint, etc.) Some perform better than others when ranking structures by their similarity (O'Boyle et al., 2016).

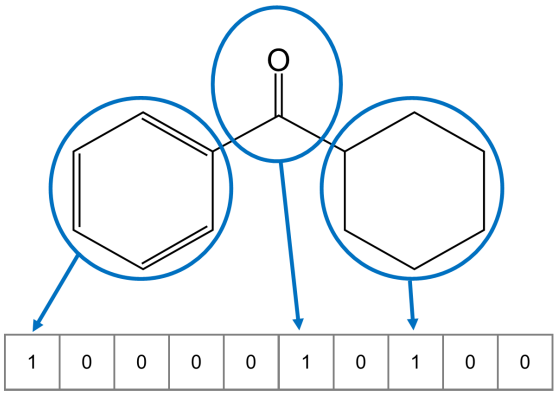


Fig. 2.5: An example of a hypothetical 10-bit molecule fingerprint. Each bit represents presence or absence of a particular structural feature of the molecule.

Subsequently, the similarity of molecules can be determined by their fingerprints comparison. The process of comparing molecular fingerprints is known as similarity search (fingerprint-based virtual screening) and has been used as an established

effective approach in drug discovery and development research (Cereto-Massagué et al., 2015).

There are multiple methods used to calculate the distance or similarity, such as Euclidean distance, Manhattan distance, distance vector, Tanimoto similarity, Cosine similarity, Dice's coefficient, or Levenshtein distance measure. Although there is a variety of similarity metrics, in a recent review Tanimoto index, Dice index, Cosine coefficient, and Soergel distance were concluded as the best options for quantifying the similarity of molecules (Bajusz et al., 2015). In this study we use the Tanimoto coefficient, as it represents one of the most widely used metrics for fingerprint comparisons.

The Tanimoto coefficient computes the degree of the similarity between two structures as the ratio of the common bits in bit vector fingerprints and is calculated as follows:

$$Tanimoto_{(A,B)} = \frac{c}{a+b-c} \tag{2.1}$$

in which a is the number of '1' bits in molecule A; b is the number of '1' bits in molecule B; and c is the number of '1' bits common in both molecules A and B (intersection). Number '1' indicates the presence of a specific structural feature for a given molecule. The range of Tanimoto values is 0–1 from the least to the most similar/identical. The figure below represents how molecules can be compared by a fingerprint similarity calculation (Fig. 2.6). The drug molecules in the figure are represented as fingerprints. Consequently, a query molecule similarity to other drug molecules in a database is calculated using Tanimoto similarity coefficient. Those molecules with a Tanimoto coefficient above a set threshold value of 0.7 are then labeled as similar to a query molecule.

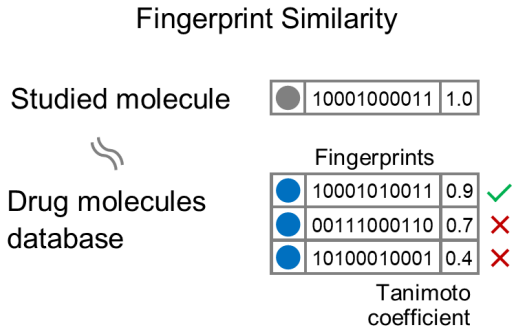


Fig. 2.6: An example of fingerprint similarity usage. The drug molecules are represented as fingerprints. Consequently, fingerprints are compared in terms of similarity using Tanimoto similarity coefficient. Those molecules with a Tanimoto coefficient above a set threshold value of 0.7 are then labeled as similar to a query molecule.

Besides 2D structural fingerprints, there are also pharmacophores representing 3D fingerprints. According to the IUPAC (International Union of Pure and Applied Chemistry) definition, a pharmacophore is an ensemble of steric and electronic features which is necessary to ensure the optimal supramolecular interaction with a specific biological target structure and to trigger (or block) its biological response (Wermuth et al., 1998) (Fig. 2.7). Pharmacophore modeling has been applied as an important tool in various approaches of computational drug discovery lately (Koščová et al., 2016). However, it has been shown that 3D shape-based approaches do not always give better results than simpler and faster similarity search approaches (Venkatraman et al., 2010). Moreover, structural information of targets is not always available.

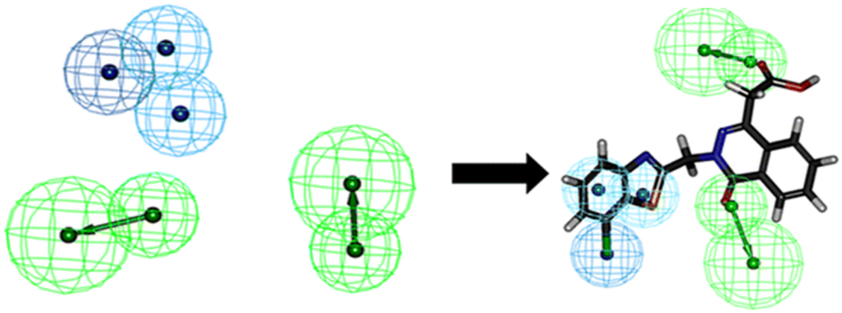


Fig. 2.7: A pharmacophore model visualization. Pharmacophore represents an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interaction of molecule with a specific biological target structure. Adapted from Sakkiah et al., 2012.

A morphine rule is a good example explaining the concept of molecular similarity. The rule states that the role of the shape of the morphine molecule referring to similarities of opioid structures is crucial in fitting exactly to the receptor active site (Myers, 2007). Molecules fulfilling the morphine rule share a set of structural features that are responsible for the same bioactivity of the molecules - mimicking the action of endorphins and relieving pain. The features include tertiary amine, quaternary carbon, a phenyl ring connected to a quaternary carbon, a two-carbon chain between a tertiary amine and a quaternary carbon (Fig. 2.8).

In the same manner, we can use similarity-based methods to calculate a measure of the degree of the overlap between a pair of molecules in other property spaces, such as the space of the side effects, indications, targets or interacting drugs. It is established from a variety of studies in which drugs with common drug targets and similar therapeutic effects are involved in similar signaling cascades and tend

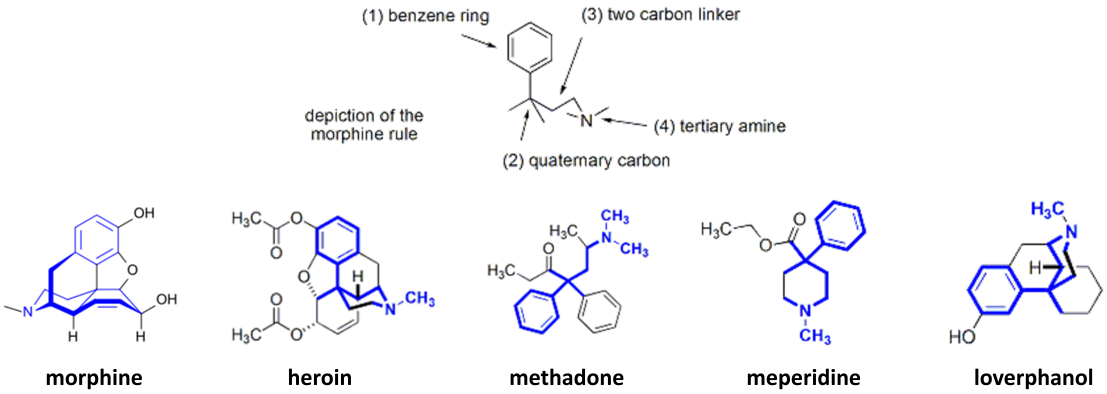


Fig. 2.8: The morphine rule. Molecules fulfilling the rule share a set of structural features that are responsible for the same bioactivity of the molecules - mimicking the action of endorphins and relieving pain. Adapted from Stevens, 2016.

to have similar side effects. Motivated by this assumption, we focus on the problem of predicting side effect prediction using similarity-based approach in our work.

Many drugs can present multi-target activity and interact with more than one therapeutic target in the human body. This drug feature is called polypharmacology. It has been shown that the most important reason why the drugs are promiscuous is the binding site similarity of their (different) targets (Haupt et al., 2013).

There are many patients who receive treatments with multiple medications. Potential drug-drug interactions can increase the risk of adverse drug reactions. The interactions of co-administered drugs can be caused by many factors affecting the ADME processes. The outcome of such interactions can result in reduced efficacy of the medication or exploited adverse reactions.

The Jaccard similarity index is one of the way that can be used for two-sets comparison. It is defined as the size of the sets intersection divided by the size of the sets union and is given as:

$$J_{(X,Y)} = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \tag{2.2}$$

in which $J_{\mathrm{X,Y}}$ is the Jaccard similarity index for sets X and Y; X is set 1; and Y is set 2. If both X and Y are empty, we define $J_{\mathrm{X,Y}} = 1$, and:

$$0 \leq J_{(X,Y)} \leq 1 \tag{2.3}$$

The Jaccard index values range from 0 to 1 corresponding to 0 to 100 % similarity. In fact, Tanimoto coefficient is a generalization of the Jaccard similarity index which is incorrectly regarded as the same sometimes.

2.4 Machine learning in drug discovery

Data mining represents a process of discovering new knowledge from large amounts of data. Machine learning methods are chosen as a common approach in the field of data mining as they are used to build models which help to discover hidden useful patterns and trends and provide valuable insights into the data. They have been successfully applied in many scientific areas and different fields of daily use such as data security, financial trading, marketing, language processing, smart car industry, or healthcare. It has been proven that such approaches have significant potential to accelerate also drug discovery attempts. For example, the techniques are helpful and widely used for drug–target interaction prediction approaches (Bagherian et al., 2020), structure-based binding affinity prediction (Ain et al., 2015) or molecular docking (Khamis et al., 2015). More applications have been reviewed elsewhere (Gertrudes et al., 2012; Lavecchia, 2015; Yang et al., 2019; Vamathevan et al., 2019).

In machine learning process, data are translated into features which are used to train a predictive model. The process consist of several fundamental steps – data collection, data training and testing in machine learning algorithm, and model (the output of machine learning algorithm) evaluation (Fig. 2.9).

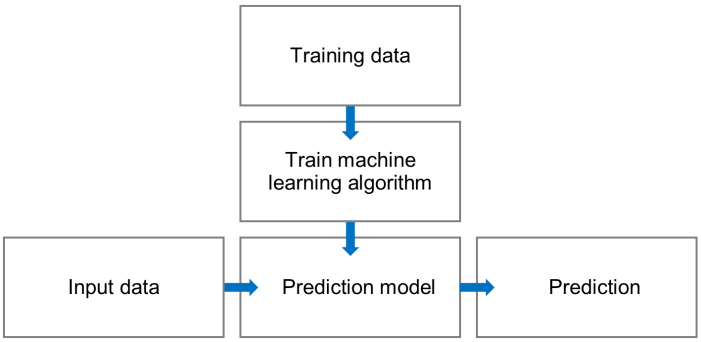


Fig. 2.9: A simplified machine learning sequence. Data preparation step includes data acquirement, data filtering and data exploratory analysis. Model performance evaluation is also a necessary part of the sequence.

We can divide machine learning techniques into two main categories – supervised and unsupervised approaches. For the supervised methods there are known outputs (targets) for input data available and the program needs learn on some example data – these data are labeled. On the other hand, for unsupervised methods there are no known or available outputs provided for input data – the data is unlabeled. Furthermore, there are semi-supervised machine learning methods which are combination of supervised and unsupervised methods – part of the inputs has known outputs and part does not (expectation–maximization algorithm).

In our work, we try to predict side effects based on known outputs, thus we apply a supervised machine learning approach. In general, the supervised tasks can be divided to classification or regression tasks. The first one predicts a category, the second one predicts a numeric value.

The goal of the classification algorithms is to classify test data (predict their category) using a model trained on collection of attributes of train data. If there are only two categories, the classification is referred to as binary classification. If there are more than two categories, it is multi-class classification. There are many different algorithms which can be used for classification tasks including decision tree, random forest, K nearest neighbors, or gradient boosted trees. Each algorithm has its own pros and cons.

The aim of our work is to predict side effect associations using a model trained on collection of data with two attribute categories, thus we apply a binary classification. Easy-to-understand and quick-to-implement algorithm called decision tree has been selected. Decision tree is a traditional supervised machine learning method, which has been commonly used in drug discovery (Blower et al., 2006; Costa et al., 2010; Bresso et al., 2013). The algorithm can be represented as a tree drawn upside down with its root at the top (Fig. 2.10). Each branch value corresponds to a possible value of an attribute defined in a decision node (Mitchell, 1997). There are several benefits of decision trees including their relatively fast construction in comparison to other methods. Moreover, they are easy to understand and nonlinear relationships between parameters do not affect the model performance.

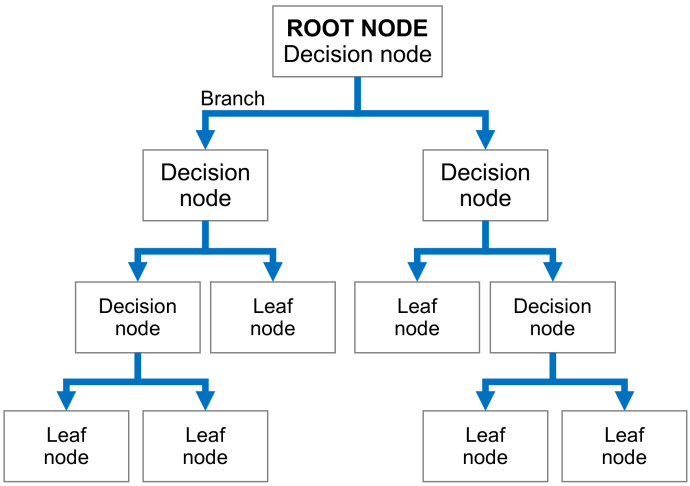


Fig. 2.10: A schematic illustration of decision tree algorithm. The tree consists of branches (observations of item), decision nodes (decisions how to split the dataset into subsets) and leave nodes (target values). The root node represents entire dataset being analyzed.

2.5 Research studies related to computational side effect prediction

In recent years various computational approaches have been proposed to predict drug side effects (Sachdev et al., 2020). In particular, there has been a growing interest in applying machine learning techniques in this field. However, despite many successful analyzes, there is currently no widely accepted standard protocol for drug side effect data curation.

The concept of drug similarities has already proven to be helpful in side effect research. Many studies have been published on utilizing chemical structures or protein targets for predicting side effects (e.g. Xie et al., 2009; Scheiber et al., 2009; Pauwels et al., 2011). In addition to them, indication-side effect relationships have been analyzed in association analysis (Zhang et al., 2013). In the study, indication predictions using side-effects information and side effect predictions using indication were compared to predictions using only chemical structures and protein targets. As a result, it was shown that there is a significant correlation between side effects and indications and the studied features have predictive power. The research was expanded and a visualization tool was developed for interactive exploration in a further study (Wang et al., 2014).

Other promising results yielded by a study which applied a combination of correlation based analysis with network-based diffusion (Atias et al., 2011). It was shown that this strategy achieves high accuracy and that a combination of different data, such as chemical structure and cell line response, can improve prediction models performance.

Another study aimed at side effect identification by applying relational machine-learning methods (Bresso et al., 2013). The study proposed an approach integrating semantic similarity measures for predicting side effect profiles using inductive learning programming which resulted in models with a higher sensitivity than decision tree models.

A robust method was proposed to predict the adverse degree of drugs using quantitative prediction models (Niu et al., 2017). In the study, drug side effect profiles were transformed to quantitative scores by summing up side effects with weights representing their importance. Consequently, the usefulness of various drug features has been evaluated in feature-based prediction models. Promising performance results were achieved by combination of three drug-related features, namely chemical substructures, targets, and treatment indications. The robustness of the method has been tested by simulation experiments on side effects with randomly assigned empirical weights.

Data on side effect association to drugs have also been applied to drug reposition-

ing studies which aim to discover new purposes for already approved and profiled drugs rather than to discover a completely new drug *de novo*. A network-based analysis was proposed as one of suitable approaches for drug repositioning (Mohd Ali et al., 2017). In the study, the drug-side effect and drug-indication networks were constructed based on drug similarities and analyzed in terms of centrality measures. The results from the work indicated that such an approach is promising for drug repositioning attempts.

To address the problem of identifying critical features for drug-side effect association prediction, predictor using multiple information integration with centered kernel alignment has been proposed (Ding et al., 2019). In the study, multiple kernels describing the information of drugs and side effect terms were analyzed. The results show that a fusion of feature spaces and a combination of different kernels by linear weighting can improve prediction performance.

There is vast number of *in silico* tools, libraries and extensions developed for specific tasks in drug discovery and development attempts. For instance, for a side effects similarity analysis, one can use DrugClust R package (Dimitri et al., 2017). The package has been developed to calculate probability scores of side effects according to the similarity of drug chemical and biological features

Machine learning prediction of new drug side effects using disease indications and structural features was studied (Khan, 2017). One of the key findings of the study suggest that integrating indications and structural features improved the side effect predictions.

It has been demonstrated that predicting side effects via machine-learning can be optimized using the right set of drug features (Seo et al., 2020). Combining various information resources such as drug-drug interactions, single nucleotide polymorphisms, chemical structures, indications, targets, and side effect anatomical hierarchy has proven to enhance side effect prediction capability in comparison to methods dependent only on chemical, indication and target features analysis.

In another recent study, it has been claimed that models based on negative sample selection strategies produce a higher performance than those without such a strategy (Liang et al., 2020). The study explored the efficiency of the strategy based on selecting negative samples in chemical-chemical interaction networks by applying random walk with a restart algorithm and concluded that negative samples are useful considerations for the side effect prediction strategies.

2.6 Integration of workflows in drug discovery

A workflow can be defined as an organized sequence of algorithms, steps or actions taken to accomplish a particular task. Among others, the key benefits of workflow usage in research include automation, easy collaboration, process control and reproducibility.

Software systems which are used to complete workflow tasks are called workflow management systems. There is a variety of tools and software available for drug discovery workflow construction (Tiwari et al., 2007). To provide a few examples, the used workflow systems include for example Chemistry Development Kit (CDK) (Steinbeck et al., 2006), Pipeline Pilot (BIOVIA, [n.d.]), ORANGE (Demšar et al., 2013), Scaffold Hunter (Schäfer et al., 2017), or KNIME Analytics Platform (Berthold et al., 2007). KNIME (Konstanz Information Miner) is an established open-source tool for interactive data exploration which integrates various components for data extraction, data processing, data mining, or interactive visual analyzes.

2.6.1 KNIME Analytics Platform

KNIME Analytics Platform (Berthold et al., 2007) is a free workflow based data mining tool originally developed by the Michael Berthold team at the University of Konstanz. This pipelining desktop client has a user-friendly, drag and drop interface for data manipulation and connecting different tools and it provides powerful features for creating various analytics workflows. Data are analyzed within KNIME workflows using pre-programmed components (basic programming units) called nodes (Fig. 2.11) that enable the user without deeper programming abilities to perform complex data analyzes. KNIME allows users to combine more than a thousand nodes in order to create new workflows graphically. The software is accessed via KNIME web portal (`https://www.knime.com`) and licensed under the GNU General Public License. The KNIME documentation can be accessed via `https://docs.knime.com`.

KNIME software is platform-independent and provides the same results on different operating systems. It is written in Java but the scripts written in the most important scripting languages (R and Python) can also be applied using the KNIME scripting integration which allows hundreds of powerful libraries to be accessed. Other strengths of KNIME Analytics Platform lie in its extensive repository of tools and external packages and compatibility with different software (Weka, Keras, Scikit-learn, etc.)

However, the key advantage of KNIME is its integration with the life science

and chemistry plugins which other similar tools lack. There is a vast collection of special bioinformatics and chemoinformatics nodes and extensions including RDKit (*RDKit: Open-source cheminformatics*, [n.d.]), Vernalis (Roughley, 2020), KNIME-CDK (Beisken et al., 2013), Enalos+ (Varsou et al., 2018), or 3d-e-chem (McGuire et al., 2017) community extensions. Furthermore, there are some commercial extensions such as extensions to CCG MOE (*Molecular Operating Environment (MOE), 2013.08*, 2017), ChemAxon (*ChemAxon*, [n.d.]), Schrodinger (*Schrödinger KNIME Extensions*, 2021) software. There is also NodePit search engine providing an exploration of KNIME node usages (available at `https://nodepit.com`).

KNIME has been used in variety of drug discovery research. For example a KNIME workflow was developed to filter target structures matching PAINS (pan-assay interference structures) (Saubern et al., 2011), there is a visualization tool called HiTSEE (High-Throughput Screening Exploration Environment) integrated to KNIME and used for analysis of large high-throughput screening data (Strobelt et al., 2012), or CheS-Mapper KNIME extension used for visual validation of QSAR (quantitative structure-activity relationship) models (Gütlein et al., 2014). There are KNIME workflows developed for investigating data from BindingDB database (Nicola et al., 2015), for predicting chemical properties with quantitative regression models (Yin et al., 2015), or for a creation of quantitative structure–property relationship models using the command line program chemalot (Lee et al., 2017). KNIME workflows were also applied for development of VSPrep tool intended for preparation of molecules for virtual screening (Gally et al., 2017), or for data mining research in medical chemistry (DiTommaso, 2017). Also, they were used to perform data curation in a research of multi-target directed ligands against Alzheimer's disease (Ambure et al., 2019), for *in silico* prediction of human oral bioavailability (Falcón-Cano et al., 2020), or for performing ligand-based *in silico* drug repurposing (Tuerkova et al., 2020). Some other life science research projects that have integrated KNIME are presented in a recent review (Fillbrunn et al., 2017).

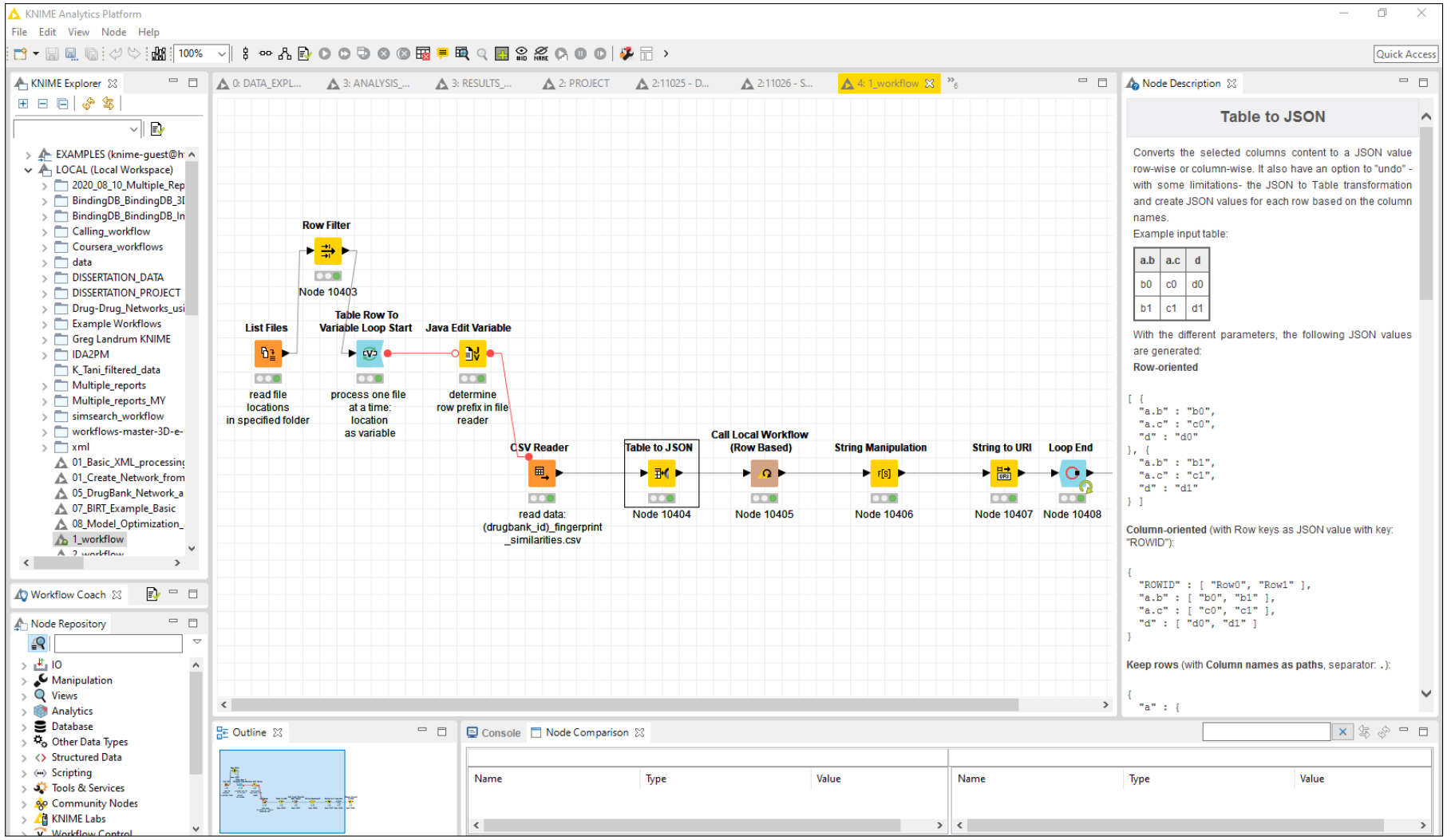


Fig. 2.11: The KNIME Analytics Platform interface (version 3.7.2). The workbench is organized into several sections, namely: KNIME Explorer, Workflow Coach, Node Repository, Workflow Editor, Outline, Console, and Node Description.

3 Methods

In this part of the thesis, we present the methodology applied in our work.

3.1 Workflow schematic representation

In our work we implemented computational techniques which integrate various drug data and result in a comprehensive set of workflows for analyzing drug–side effect associations. All datasets were preprocessed and further analyzed via a sequence of specific KNIME nodes including integrated R scripts. The nodes were involved in applying selected analysis techniques. The methodology steps of our work are illustrated below (Fig. 3.1).

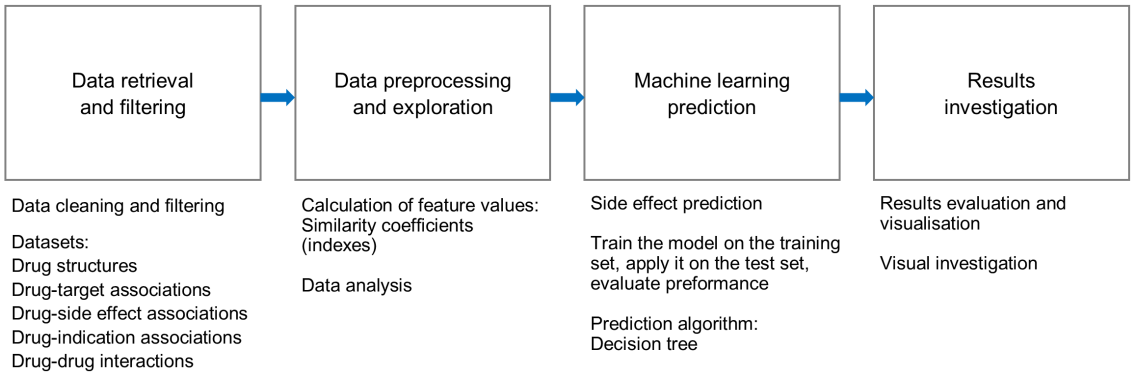


Fig. 3.1: A visual description of the methodology steps. The set of the workflows starts with data deployment and preprocessing. This is followed by a workflow intended for data exploration and visualization. The next workflows deal with machine-learning model construction and predicting side effects. In the final part, the results of the predictions are evaluated and visually investigated.

The set of the workflows is available as a part of the additional files. It can be used as a single pipeline, or as multiple stand-alone workflows to gather and prepare data, to calculate similarities, to run classification models, to inspect model performances and to evaluate predictions, respectively.

Each of the following subchapters focuses on a different aspect of the methods used, however, the basic idea of our work is as follows: combined similarity characteristics of the drugs are used as features for predicting potential side effects. The goal of the set of the workflows is to investigate if a specific drug molecule would have a specific side effect according to a prediction based on multiple similarity metrics to other drug molecules and their association with a given side effect. We calculated the similarity between drugs based on their chemical structure and association with

specific side effects, targets, indications, and interacting drugs. The assumption is that two drug molecules with similar structure, side effects, target proteins, indications and interacting drugs will also be more similar in terms of association to a query side effect.

We constructed the workflows with KNIME software (version 3.5.2., released on April 18, 2019) that is available at `https:/www.knime.com`. The applied nodes and their function in the workflow are briefly described below. In addition, several own scripts written in R language were integrated into the workflows. R is a programming language and free software environment used mainly for statistical computing and graphics (Team, 2008). All the analysis was performed on a stationary computer with Intel(R) Core(TM) i5 CPU and 16.0 GB memory.

3.2 Data retrieval and filtering

The amount and quality of data are important factors for machine learning models to learn the tasks. Here we describe the data used in this work and the filters applied on them before the performed analysis.

3.2.1 The data sources

Our data was acquired and imported to KNIME software from the openly available sources listed in the following table (Tab. 3.1). The datasets consist mainly of information about drugs, side effects, targets, indications, and interacting drugs. All the files are available in the attachments or via referred websites. The scripts used for data collection is available in the attachment as well. Datasets statistics are summarized in the Results and discussion section (Chapter 4).

Information about drugs, targets and interacting drugs were collected from DrugBank database (version 5.1.7, released on July 2, 2020) which is freely accessible in .xml format at `https://go.drugbank.com/releases/latest`. In order to be able to process data from DrugBank xml file also in KNIME, we used *Mohammed-FCIS/dbdataset* data package to retrieve the data as data frames parsed by dbparser R package (version 1.2.0, released on August 8, 2020) (M. et al., 2020). The parsed DrugBank dataset is available after download from `https://github.com/interstellar-Consultation-Services/dbdataset`, installing the package by `devtools::install_github('MohammedFCIS/dbdataset')` command and loading it by `library(dbdataset)` command in R environment. The SIDER database (version 4.1, released on October 21, 2015) was used as a source for related information on drug side effects and indications. This version uses side effect names

from MedDRA dictionary. The files are available at `http://sideeffects.embl.de/download`.

Tab. 3.1: The data sources. Various data was acquired from DrugBank and SIDER databases.

Database	**Source file**	**Data type**
DrugBank	DrugBank_structure_links.csv	drug structures
	R library 'MohammedFCIS/dbdataset': data(Targets_Drug)	drug targets
	R library 'MohammedFCIS/dbdataset': data(Interactions_Drug)	interacting drugs
SIDER	meddra_freq.tsv	drug side effects and their frequencies
	meddra_all_indications.tsv	drug indications
	drugs_atc.tsv	drug ATCs

After retrieval, some of the data needed to be filtered out to meet the necessary requirements for further analysis. The reader can review data overall filtering process in the flowcharts below. The workflow nodes access the DrugBank database to get information about structures, targets, indications, and interaction drugs of desired drug molecules, and SIDER database to get information about side effect frequencies and indications (Fig. 3.4). The filtering process is further described in the following subchapters. The intersection drug set represents the set of all drug IDs which are members of all of the filtered datasets (Fig. 3.2). Those drugs are referred to as examined drugs in the following text.

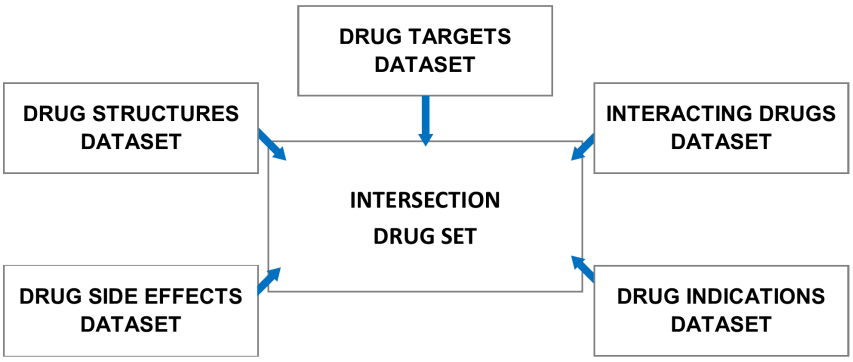


Fig. 3.2: The dataset filtering flowchart. The intersection drug set corresponds to set of examined drugs.

3.2.2 Drug molecule chemical structures

There are several widely used chemical file formats of drug compounds. In order to deal with molecules in this study, we worked with a 2D string representation called simplified molecular input line entry system (SMILES). SMILES is a line notation which follows a few syntax rules to encode molecular structures as a linear ASCII string, see the example below (Fig. 3.3). The main benefit of the SMILES format is that it has low storage requirements, making it ideal for storing large molecule datasets.

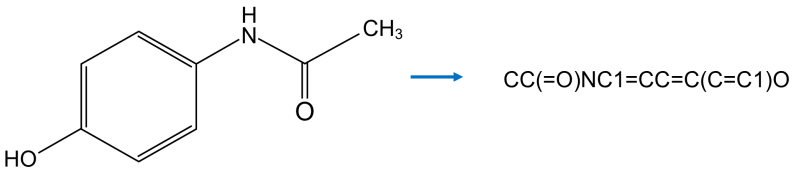


Fig. 3.3: An example drug molecular representation in SMILES format - a line notation that follows few syntax rules to encode molecular structures as a linear ASCII string.

Drug structure data of drugs were downloaded as *DrugBank_structure_links.csv* file available at `https://go.drugbank.com/releases/latest#structures`. The file includes structure information in InChI, InChI Key, and SMILES format and identifiers for other drug-structure resources. Drug SMILES representations were used to generate fingerprints in a later step. For our analysis we only used drugs which were labeled as approved small molecules and had SMILES data available (Fig. 3.4).

3.2.3 Drug side effects

The SIDER database was used as a source for drug side effects (version 4.1, released on October 21, 2015). We obtained side effects including frequency data from *meddra_freq.tsv* file and the dataset processing filter was applied in the following way (Fig. 3.4).

At first, we only considered drugs which had not been used as placebos in safety studies. Secondly, we categorized the side effects according to CIOMS standard recommendations (CIOMS, 1995) on side effect frequency classification as very common, common, uncommon, rare or very rare (Tab. 2.3) according to the following rules: (1) an upper bound on the frequency value was applied for side effect frequency classification, (2) if there were more than one frequency value, a minimum value was applied for side effect frequency classification.

Next, we excluded all lowest level term side effects according to the MedDRA concept type. Only preferred terms of side effects remained.

Furthermore, only side effects with frequency data available were considered for our study. We filtered out only those side effects, which were categorized as 'very common' (frequency of $\geq$ 10%) and 'common' (frequency of $\geq$ 1% and <10%) according to the CIOMS classification (Tab. 2.3). The side effects in other categories (very rare, rare, uncommon, common) were not considered in our analysis. Consequently, drugs with no 'very common' side effects were excluded from the analysis.

Entries without necessary side effect information available were avoided and cases in which side effects were found with both positive and negative association, we kept the positive one.

Consequently, we mapped drug identifiers from DrugBank (DrugBank ID) to SIDER (STITCH ID flat) via the common ATC codes (Anatomical Therapeutic Chemical) available from *drugs_atc.tsv* file. The drugs with ATCs (STITCH ID flat) mapped to more than one DrugBank IDs were excluded from the analysis. Only those side effects which are associated to approved small molecules drugs and have SMILES data available were included while drugs with no known side effect association information were not.

3.2.4 Drug indications

Drug indication represents drug association to a disease or condition. The *meddra_all_indications.tsv* file (obtained from SIDER database) was used as a source for drug indication dataset (Fig. 3.4). We obtained a set of indications associated to drugs filtered in a previous step. A further dataset filter was applied in the following way. At first, we excluded rows with missing *UMLS_id_MedDRA_indication* information. Indication IDs obtained from label (*UMLS_id_label_indication*) were not considered. Next, we excluded all lowest level term indications according to the MedDRA concept type. Only preferred terms of indications remained. Finally, we included only indications associated to the examined drugs filtered as described in previous sections in the dataset.

3.2.5 Drug targets

DrugBank R library *MohammedFCIS/dbdataset* was used as a source for drug targets dataset. The dataset can be obtained after loading the library and running the following command `data(Targets_Drug)` (Fig. 3.4). We filtered all targets associated to the examined drugs filtered as described in previous sections.

3.2.6 Interacting drugs

DrugBank R library *MohammedFCIS/dbdataset* was used as a source for interacting drugs dataset. The dataset can be obtained after loading the library and running the following command `data(Interactions_Drug)` (Fig. 3.4). We filtered all interacting drugs associated to the examined drugs filtered as described in previous sections.

3.3 The similarity metrics calculation

This chapter describes the calculating of similarity metrics used in this work. Additionally, we explain the data exploration process which was performed before feeding the similarity data into machine learning models.

3.3.1 The structure fingerprint similarity computation

At first, 2D fingerprint similarity was employed for structure similarity description as follows. Multiple fingerprints were generated for all examined drugs. In total, eight different most common fingerprints were used for the similarity calculation between the query drug and the remaining dataset drugs based on the chemical structure, namely: Morgan fingerprint, FeatMorgan fingerprint, AtomPair fingerprint, Torsion fingerprint, RDKit fingerprint, Avalon fingerprint, Layered fingerprint, and MACCS fingerprint as described in the table below (Tab. 3.2).

Tab. 3.2: Calculated drug molecule structural fingerprints description (adapted from RDKit documentation (Landrum, [n.d.])).

Fingerprint name	**Description**
RDKit	A daylight-like topological fingerprint based on hashing molecular subgraphs
FeatMorgan	A FCFP-like (Functional Class Fingerprint) circular fingerprint based on the Morgan algorithm and feature invariants
AtomPair	An atom-pair fingerprint (Carhart et al., 1985)
Torsion	A topological-torsion fingerprint (Nilakantan et al., 1987)
Avalon	Avalon toolkit fingerprint (`https://sourceforge.net/p/avalontoolkit`)
Layered	An experimental substructure-matching fingerprint
MACCS	A SMARTS-based (SMiles ARbitrary Target Specification) implementation of the 166 public MACCS keys (Molecular ACCess System) (Durant et al., 2002)
Pattern	A topological fingerprint optimized for substructure screening

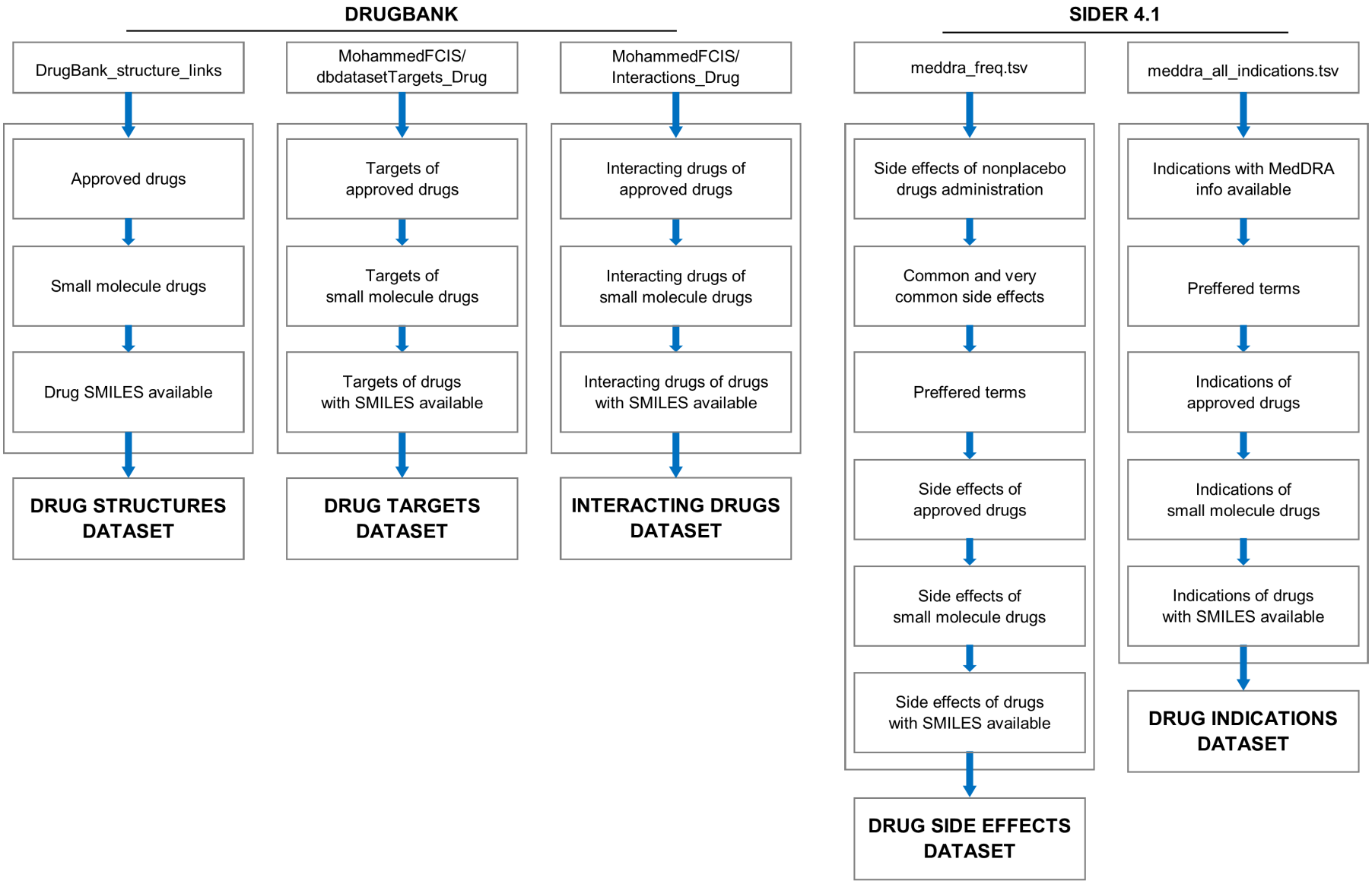


Fig. 3.4: The flowcharts to obtain and filter data from DrugBank and SIDER

Consequently, the drug similarity was computed. The aim of similarity calculation was to compare the similarity of a query (reference) molecule fingerprint to the database molecule fingerprints. The similarity of molecules was computed by calculating the maximum Tanimoto similarity coefficient based on the CDK toolkit (Beisken et al., 2013) for all structures in provided datasets. The Tanimoto calculation was performed based on the aggregation method.

The Tanimoto similarity coefficient of drug molecule structures is given by:

$$Tanimoto_{(A,B)} = \frac{c}{a+b-c} \tag{3.1}$$

in which a is the number of '1' bits in molecule A; b is the number of '1' bits in molecule B; and c is the number of '1' bits common in both molecules A and B. Number '1' indicates the presence of a structural feature for a given molecule. In total there were 8 different Tanimoto coefficients calculated (for each type of generated structural fingerprints).

3.3.2 The Jaccard similarity index calculation for the shared side effects, indications, targets, and interacting drugs

The figure below depicts the process of the similarity calculation (Fig. 3.5). The sets of side effects, indications, targets, and interacting drugs of each examined drug in the datasets were converted to bit-vector fingerprints. Following this, the Jaccard similarity index was used as a method to calculate similarity for the query drug molecule in terms of shared side effects, indications, targets, and interacting drugs association. All types of the indexes were calculated for each drug–drug pair. The drug–drug pairs rows with 0 similarity were excluded from the dataset, as well as duplicated drug–drug pair combinations. The Jaccard similarity index for shared features is given by:

$$J_{(A,B)} = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3.2}$$

in which J is the Jaccard similarity index for a specified feature; A is drug 1 set of features; and B is drug 2 set of features. As all the drugs in the dataset had at least 1 feature, none of the sets was empty.

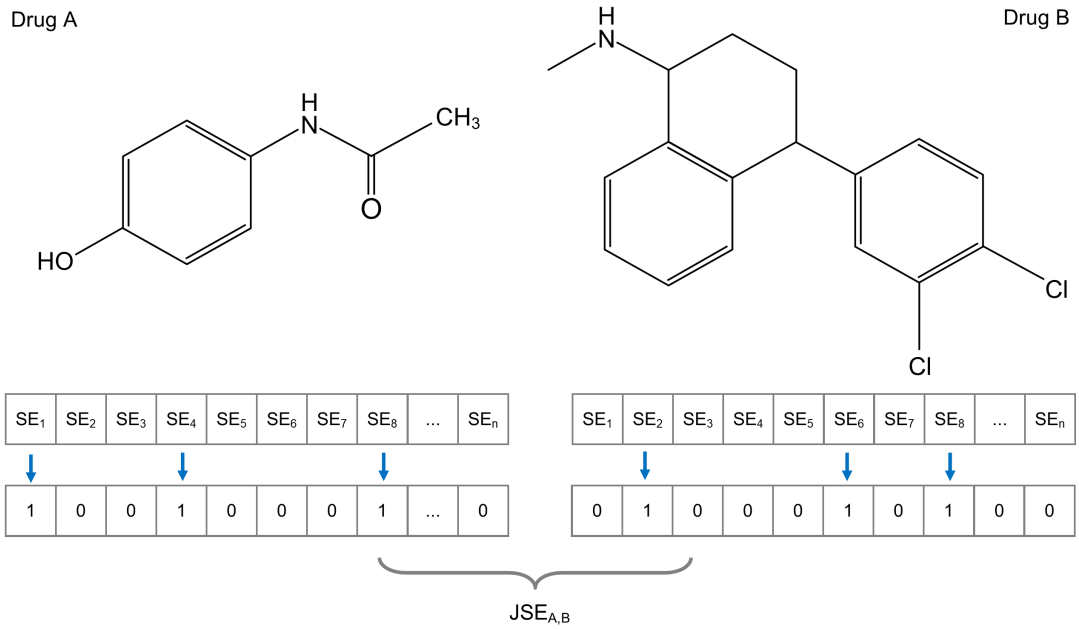


Fig. 3.5: The featured fingerprints and the Jaccard similarity index calculation (J). Presence of specific features are compared between two drugs. In this example, side effects associations (SE) are compared.

3.4 The datasets for the analysis construction and exploration

The Tanimoto structural fingerprint similarity coefficients, the Jaccard feature similarity indexes and the side effect association were merged together to compile datasets for analysis. Each of the resulting tables with the similarity measures serves as a required input data of attribute matrix describing the examined drugs. As all fingerprint similarities were numerical values in range of [0, 1], there was no need to implement a normalization step on the data. Several erroneous molecules were excluded. In most cases the cause of errors was an inability to generate a SMILES structure.

As a result of the similarity calculation step, each examined drug in each final dataset was represented by a 12-dimensional profile whose elements encode for the structure fingerprint similarity coefficient or the Jaccard similarity index for side effects, indications, targets, and interacting drug similarities to a query drug molecule. The studied side effect association was added in the last column as a target attribute. It was a nominal value corresponding to a positive or a negative association to a given side effect. The models were constructed for predicting the top 10 most prevalent side effects in the filtered dataset. They are listed in a table below (Tab. 3.3).

Tab. 3.3: The studied side effect predictions

UMLS ID	MedDRA info 3
C0000737	Abdominal pain / Gastrointestinal pain
C0004604	Back pain
C0009806	Constipation
C0011991	Diarrhoea
C0012833	Dizziness
C0013395	Dyspepsia
C0015672	Fatigue / Asthenia
C0018681	Headache
C0027497	Nausea
C0042963	Vomiting

3.4.1 Features distribution

In order to analyze feature distribution, we divided the prepared dataset into two parts, with and without a given side effect association respectively. The data distributions of all features were studied using a comparison of box-and-whisker plots leading to a conclusion if there tends to be a difference between both groups (the group of examined drugs with positive vs. the group of examined drugs with negative side effect associations).

The box-and-whisker plot is a diagram used to summarize a set of numerical data by visualizing it through their quartiles and displaying the shape of the distribution (possible skewness), central value, and variability (spread) (Tukey, 1977). Nowadays, box-and-whisker plots are one of the most frequently used statistical graphic. They are particularly useful for distributions comparisons between datasets. The distance between medians as a percentage difference of the overall visible spread was calculated as follows:

$$P = \frac{DBM}{OVS} \times 100 \tag{3.3}$$

in which P is the percentage difference, DBM is the difference between medians corresponding to the difference between medians of the both groups, OVS is the overal visible spread corresponding to the difference between the higher upper quartile and the lower lower quartile. For more details refer to the figure below (Fig. 3.6).

After obtaining the results of the feature distribution analysis, additional filtering was applied to get features with a greater value difference between the groups.

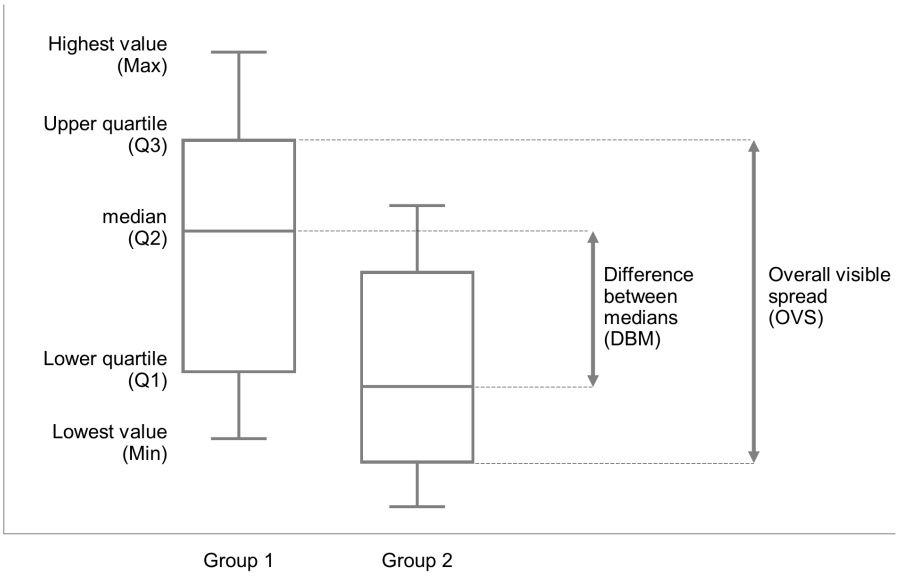


Fig. 3.6: The robust statistical parameters displayed by the box-and-whisker plots. Difference between medians is the difference between medians of the both groups, overall visible spread represents the difference between the higher upper quartile and the lower lower quartile of the bothe groups.

The filtering step was performed based on the feature robust statistical parameters (median, lower quartile, upper quartile) of both groups. In order to make the percentage difference larger and get datasets with more predictive power, we only considered drugs according to the following filtering function (Fig. 3.7) resulting in 'selection datasets'. In total 5 selection datasets have been prepared this way. The prepared datasets for analysis were fed in models in the following steps.

3.4.2 Feature selection (dimensionality reduction)

Before proceeding with further data analysis, we implemented a dimensionality reduction technique (the number of input features reduction) in order to get alternative datasets and affect the workflow speed. The benefits of the feature selection step include overfitting (model fitting to noise in training data) reduction, performance improvement and training time reduction. There are a variety of tools which can be employed to aid feature selection. The selected statistical measures calculated for this task included variance, standard deviation, coefficient of variation and correlation in our work.

The variance is a measure of how far each value in the dataset is from the mean. Hence, low variance indicates that data points are generally similar and do not vary widely from the mean. Data columns with low variance (almost constant value) could be considered for removal based on the assumption that low variance features

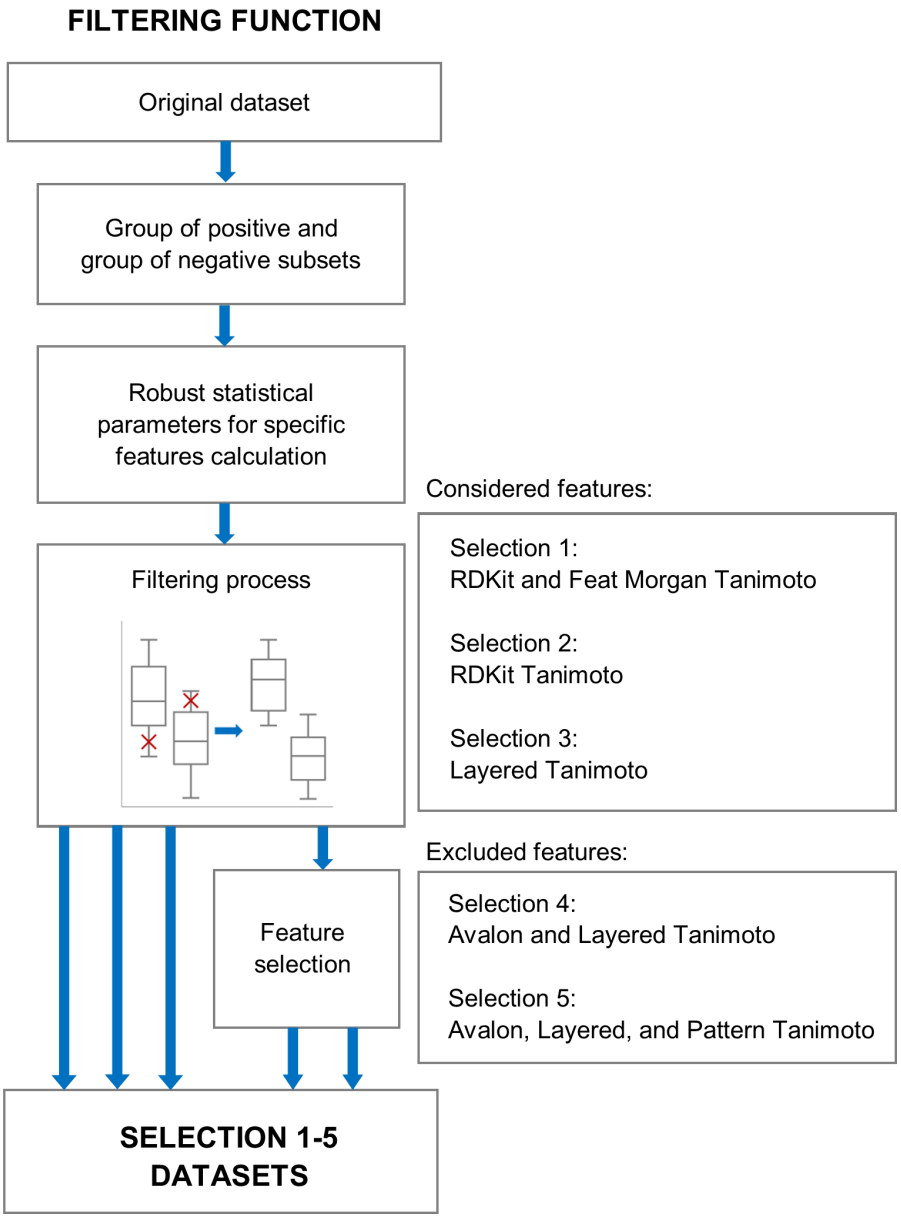


Fig. 3.7: The applied filtering function diagram. A dataset was split into a group of drugs with positive association (positives) and a group of drugs without association to a specific side effect (negatives). Median, lower quartile, upper quartile were calculated for specific feature values of both groups. If median of the group of positives was lower than median of the group of negatives, the group of positives was subset to rows with values less than upper quartile of the group and the group of negatives was subset to rows with values greater than group lower quartile of the group. Otherwise the groups were split in an opposite way. In addition, two more selection datasets have been prepared by excluding some of the features with high correlation to other features.

hold no predictive power and do not meaningfully contribute to the model's predictive capability. The sample variance was calculated for all features in the datasets as follows:

$$\textit{Sample variance} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3.4}$$

in which x is the value; $\bar{x}$ is the sample mean; N is the sample size.

Another consideration to discard a feature is based on the standard deviation calculation and the assumption that features with a standard deviation equal to zero do not vary and would have no effect on the model performance. The sample standard deviation was calculated for all features in the datasets as a square root of the sample variance as follows:

$$\textit{Sample standard deviation} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3.5}$$

in which x is the value; $\bar{x}$ is the sample mean; N is the sample size.

The coefficient of variation (also called the relative standard deviation) is a standardized measure of the dispersion of a probability distribution or a frequency distribution. The sample coefficient of variation was calculated for all features in the datasets as follows:

$$\textit{Sample coefficient of variation} = \frac{s}{\bar{x}} \tag{3.6}$$

in which s is the sample standard deviation; and $\bar{x}$ is the sample mean. The range of the coefficient of variation is between 0 and 1. Optionally it can be expressed as a percentage when multiplied by 100.

To explore feature correlation, we calculated the Pearson product-moment correlation coefficient, which is a typical approach for measuring the significance of the association between two normally distributed variables. The sample Pearson correlation coefficient was calculated according to the following formula:

$$\textit{Pearson correlation coefficient} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3.7}$$

in which x_i and y_i, are the individual sample values indexed with i; $\bar{x}$ and $\bar{y}$ are the sample means; N is the sample size. The range of the correlation coefficient is between -1 and 1. If the coefficient is equal to 0, there is no correlation between the two variables. The closer to -1 or 1 the value is, the stronger the correlation between the two variables indicates either negative or positive.

3.5 The applied machine-learning algorithm and the evaluation metrics

In this subchapter more details on the applied machine-learning techniques and the evaluation metrics are described.

3.5.1 The decision tree

A decision tree was used as a machine-learning classification algorithm. Our task is a binary classification problem in which each examined drug molecule in the datasets is considered to either be associated with a specific side effect (labelled positive) or not (labelled negative). The workflow provides an individual classifier for each side effect prediction.

The Gini index was selected as quality measure according to which the split point in the decision tree was calculated. It represents a measure of the dataset impurity which can be defined as a probability of an attribute value to be misclassified. The Gini index is given as follows:

$$Gini\ index = 1 - \sum_{i=1}^{n} (P_i)^2 \tag{3.8}$$

in which P_i is the relative frequency of class i. The range of the Gini index is between 0 and 1. The value of 0 indicates the purity of the classification. The main advantage of the index is its simple calculation requiring only the distribution (Shafer et al., 1996).

In machine-learning modeling it is necessary to have a dataset partitioned into a training set and a test set. The data of a training set are used to train the model, while test set data are required to evaluate model performance. We applied a relative split of 80% for the train part, the remainig 20% serves for model testing. A random seed 123456789 was used in random sampling to partition data (Fig. 3.8).

3.5.2 K-fold cross validation

The K-fold cross validation approach was implemented for a better model evaluation. In the first step of the cross validation process, the dataset is split into k parts (folds). Those parts are analyzed in each iteration as depicted in the figure below (Fig. 3.9). Each fold is used once for testing, remaining $(k-1)$ folds are used for training which means there is a partial overlap in each iteration.

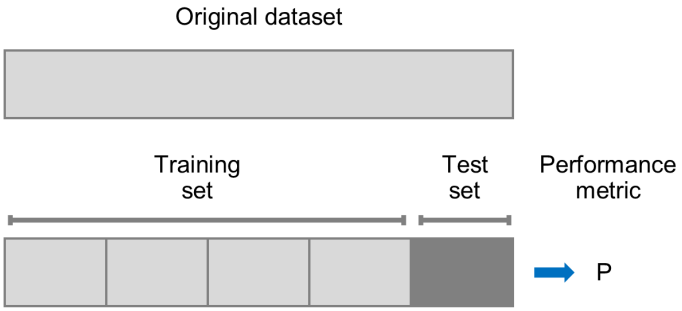


Fig. 3.8: A schematic illustration of the dataset partitioning. 80% of the dataset is intended for model training, the remainig 20% serves for model testing.

This process results in k number of models and the final cross validation performance represents an arithmetic mean of the performance of all these models. It is calculated as follows:

$$P = \frac{1}{k} \sum_{i=1}^{k} P_i \tag{3.9}$$

in which P is the cross validation performance; and P_i is the iteration i model performance.

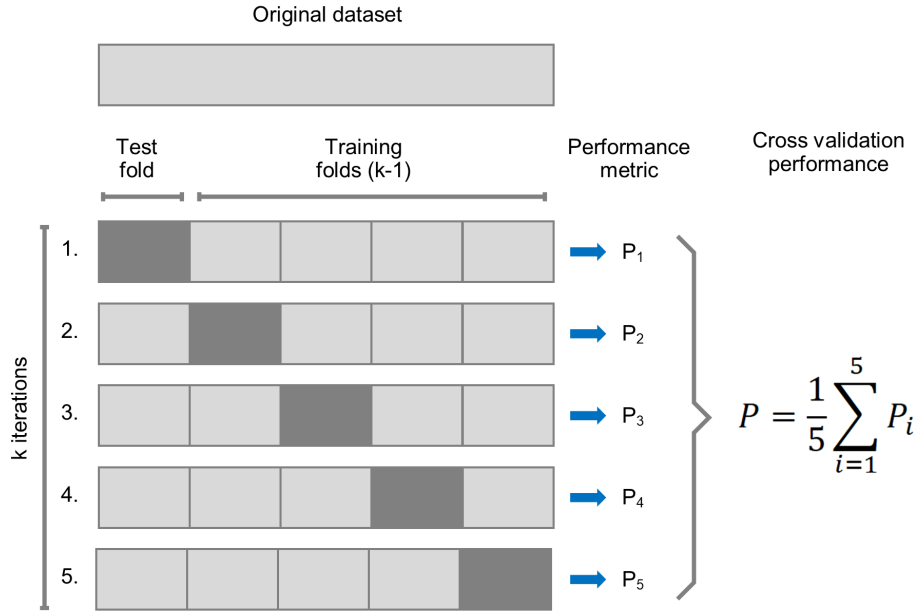


Fig. 3.9: A schematic illustration of 5-fold cross validation process. The cross validation performance represents arithmetic mean of performance of all models trained on different (overlapping) parts of the original dataset within 5 iterations.

The main benefit of a cross validation technique is that it provides more training data, and the measured estimation of model performance is more accurate. Further-

more, it is used to prevent overfitting and reduce variability and pessimistic bias. In machine learning practices, a 10-fold cross validation is very common to train and test classifiers and it was also employed in our study.

3.5.3 Prediction statistics – the performance measurement

Model evaluation is an indispensable part of machine-learning based predictions, therefore choosing a proper evaluation metric is crucial. To assess the performance of the methods applied in our work, at first, the following performance metrics for each prediction model were defined: true positive (TP), false negative (FN), false positive (FP) and true negative (TN) values. A true positive (TP) result refers to a correctly predicted positive outcome (known drug–side effect association), a true negative (TN) indicates a correctly predicted negative outcome (known lack of drug–side effect association). A false positive (FP) result refers to an incorrectly predicted positive outcome (incorrectly predicted drug–side effect association), a false negative (FN) indicates an incorrectly predicted negative outcome (incorrectly predicted lack of drug–side effect association). The above mentioned values can be summed up in a confusion matrix (also called error matrix) (Fig. 3.4). It is a table commonly used to represent the accuracy of the predictor. True positive, true negative, false positive and false negative values were used for other performance metrics calculation as described further.

Tab. 3.4: The confusion matrix for binary classification problems

	Observed positive	Observed negative
Predicted positive	Number of true positives (TP)	Number of false positives (FP)
Predicted negative	Number of false negatives (FN)	Number of true negatives (TN)

Recall (also called sensitivity in binary classification or the true positive rate) is a proportion of the true positives. In other words, recall indicates how well the model predicts the true positives category. The recall score is calculated as follows:

$$\textit{Recall (sensitivity, true positive rate)} = \frac{TP}{TP + FN} \tag{3.10}$$

in which TP is the number of the true positive outcomes; and FN is the number of the false negative outcomes.

Specificity (also called the true negative rate) indicates how well the model predicts the true negatives category. The specificity score is calculated as follows:

$$\textit{Specificity (true negative rate)} = \frac{TN}{TN + FP} \tag{3.11}$$

in which TN is the number of the true negative outcomes; and FP is the number of the false positive outcomes.

Precision (also called the positive predictive value) measures the number of true positives divided by the total number of positive outcomes. The precision score is calculated as follows:

$$\textit{Precision (positive predictive value)} = \frac{TP}{TP + FP} \tag{3.12}$$

in which TP is the number of the true positive outcomes; and FP is the number of the false positive outcomes.

Accuracy is a commonly used metric of binary classifiers which measures number of correctly classified predictions from all predictions made. The accuracy score is calculated as follows:

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.13}$$

in which TP is the number of the true positive outcomes; FP is the number of the false positive outcomes; TN is the number of the true negative outcomes; and FN is the number of the false negative outcomes.

In addition, the performance of various methods can be evaluated via a receiver operating characteristic (ROC) curve or rather the area under it. ROC is a graph in which the true positive rate (TPR) (the number of correctly classified positives to the total number of positives) is plotted on the y-axis and the false positive rate (FPR) (the number of incorrectly classified negatives to the total number of negatives) is plotted on x-axis (Fig. 3.10).

$$\textit{False positive rate} = \frac{FP}{FP + TN} \tag{3.14}$$

The area under the receiver operating characteristic curve (AUC-ROC) has been widely used in previous studies as a well-established model performance measure.

The AUC-ROC score indicates the performance of the classification model. If the AUC-ROC is equal to 1, the classifier is perfect as it is able to correctly distinguish between the positive and the negative class in 100%. The AUC-ROC score

near to the 1 indicates an excellent performance of the model. The random classifier has an AUC-ROC score of 0.5 which means such a model fails and has a bad measure of separability. A model with an AUC score near to the 0 is regarded as poor. The AUC-ROC of multiple models allows for a quick visual comparison of their performances. The higher the model ROC curve is, the better the model it indicates.

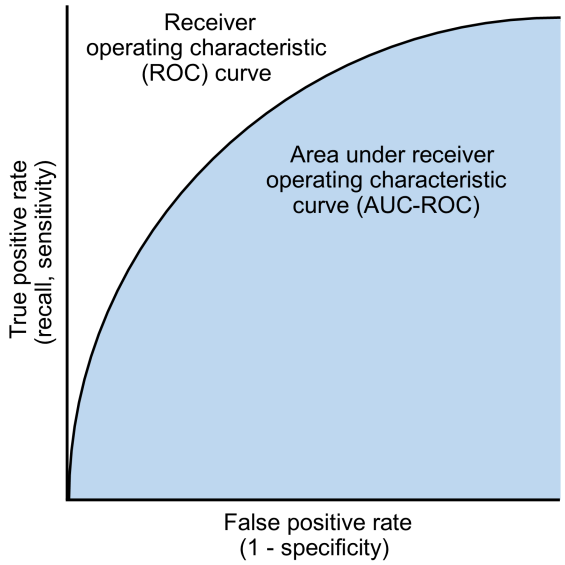


Fig. 3.10: Example of a receiver operating characteristic (ROC) curve. ROC curve plots true positive rate (the number of correctly classified positives to total number of positives) vs. false positive rate (the number of incorrectly classified negatives to the total number of negatives), in other words it plots sensitivity (recall) vs. (1-specificity). An area under ROC curve represents the probability of a correct prediction.

However, in classification problems with a large imbalanced class distribution, classification accuracy alone cannot be selected as a trusted measure due to an accuracy paradox referring to an issue in which a classifier is biased towards the majority class. Therefore, additional measures are required to evaluate model performance. As the class of negative side effect association occurs more often than the class of positive side effect association in our datasets, the datasets can be regarded as imbalanced. In the case of evaluating binary classifiers on an imbalanced data precision-recall (PR) curve analysis is more informative than an AUC-ROC curve analysis (Saito et al., 2015). PR curve is a graph in which precision (the number of true positives divided by the total number of true positives and false positives) is plotted on the y-axis and recall (the number of true positives divided by the total number of true positives and false negatives) is plotted on x-axis (Fig. 3.11). PR curve con-

nects all precision-recall points of a classifier. The AUC in the precision-recall space is called an area under the PR curve (AUC-PR).

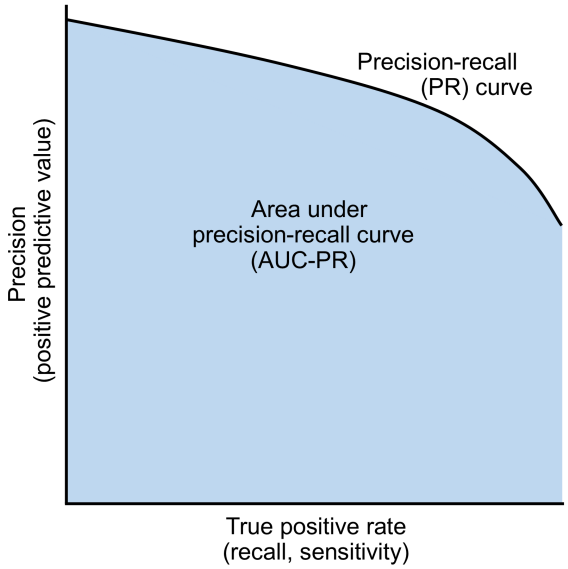


Fig. 3.11: Example of a precision-recall (PR) curve. PR curve plots the false positive rate (the number of incorrectly classified negatives to the total number of negatives) vs. true positive rate (the number of correctly classified positives to the total number of positives), in other words it plots precision vs. sensitivity (recall).

The AUC can be calculated by several methods. In our analysis, a function provided in the PRROC R package (Grau et al., 2015) was applied to calculate both the AUC-ROC and the AUC-PR values. AUC-PR was calculated using the interpolation of Davis and Goadrich (Davis et al., 2006).

The changes in all calculated performance metrics were compared for models fed with different input data. The calculated metrics are displayed using box-and-whisker plots in the following Results section.

4 Results and discussion

This chapter is intended to present the results of our work. We describe the set of developed workflows, their application on real data and the results obtained.

4.1 The details of the designed workflow usage

The final set of designed workflows consists of three main sections (Fig. 4.1, Fig. 4.22 and Fig 4.34. Their major purpose is to predict side effects on the basis of drugs similarity data. The first workflow section is intended for preparing and evaluating datasets. The prepared data are used in the second section in the machine learning models. The aim of the third section is to evaluate the results of this prediction. As a clear and meaningful visualization is a key aspect of data understanding and analysis, various plots generation is an integral part of the designed workflows. Each of the sections is described in further detail in the next subchapters. This is immediately followed by applying the corresponding parts of the workflow set to our analyzed datasets. All workflow descriptions are accompanied by figures.

The proposed set of workflows was exported to KNIME Archive File named 'DISSERTATION_PROJECT.knar' and is available as an additional file. The prerequisites to use it include having the KNIME Analytical Platform installed. The usage is platform independent and the workflows run within the KNIME Analytics Platform (KNIME version 3.7.2 or above). Other requirements include having Java 1.8 or higher installed. The KNIME Extensions mandatory for the workflows usage are listed in the table below (Tab. 4.1).

Tab. 4.1: The KNIME Extensions required for the proposed set of workflows usage. They can be installed from the KNIME Analytics Platform update site.

KNIME extension	**Info**
Erl Wood KNIME Open Source Cheminformatics	`https://www.knime.com/community/erlwood`
KNIME Base Chemistry Types & Nodes	`https://hub.knime.com/knime/extensions/org.knime.features.chem.types/latest`
KNIME Interactive R Statistics Integration	`https://www.knime.com/community/hcs-tools`
KNIME HCS Tools	`https://hub.knime.com/knime/extensions/org.knime.features.r/latest`
RDKit KNIME Integration	`https://www.knime.com/rdkit`

There are three states of KNIME nodes: Idle, Configured, Executed. To change the KNIME node configuration, right-click on the node and select 'Configure' option. To execute the KNIME node, right-click and select 'Execute' option. The green colour indicates that the node has been successfully executed. The entire workflow can be executed by selecting the 'Execute All' button on the top panel. Some of the workflow nodes have been collapsed into the metanodes containing sub-workflows to keep the pipeline tidy and easy to understand. Metanodes can be opened by double clicking in a separate tab or expanded to the original node sequence. Additionally, there are several annotations created for a better understanding.

4.2 Data retrieval and filtering

The first part of the set of workflows deals with data retrieval and filtering and the subsequent similarity metrics calculation (Fig. 4.1). The necessary datasets can be retrieved from the 'MohammedFCIS/dbdataset' library via integrated R scripts in appropriate 'R Source (Table)' KNIME nodes or imported into the corresponding workflows directly from prepared .csv files via 'CSV Reader' KNIME nodes (Fig. 4.3, 4.4 and 4.5). All files are listed in the appendix and are available and ready to use in the attachment. After data importing the workflow performs set operations required for filtering. In addition, the workflows enable the user to perform data exploration steps to better understand the nature of the analyzed datasets before the analysis itself. The resulting visualization outcomes include box-and-whisker plots and histograms. In the following subchapters you can see the results of the exploratory analysis of our datasets.

4.2.1 Drug molecule dataset after filtering

The total number of unique drugs after filtering (small molecules, approved drugs, SMILES structure available, ATC code available) was 1,898 (Tab. 4.2) (Fig. 4.2). Many of them are associated to more than one ATC code.

Tab. 4.2: The drug dataset information after filtering

Characteristics	**After filtering**
Number of drugs (unique drugbank_id terms)	1,898
Number of drugs (unique ATC terms)	1,806

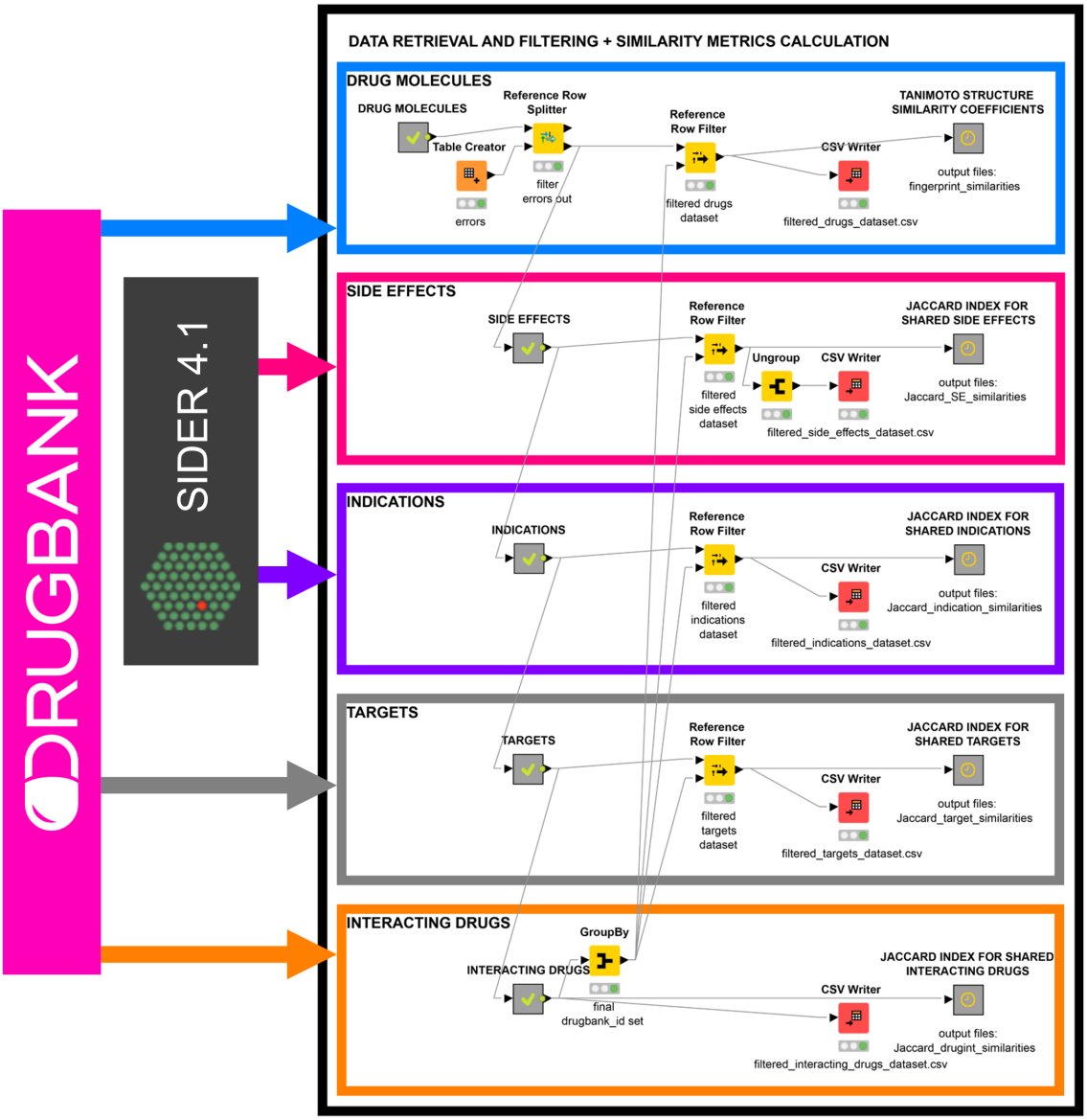


Fig. 4.1: Workflow overview - part I: a data retrieval, a data filtering and a similarity metrics calculation.

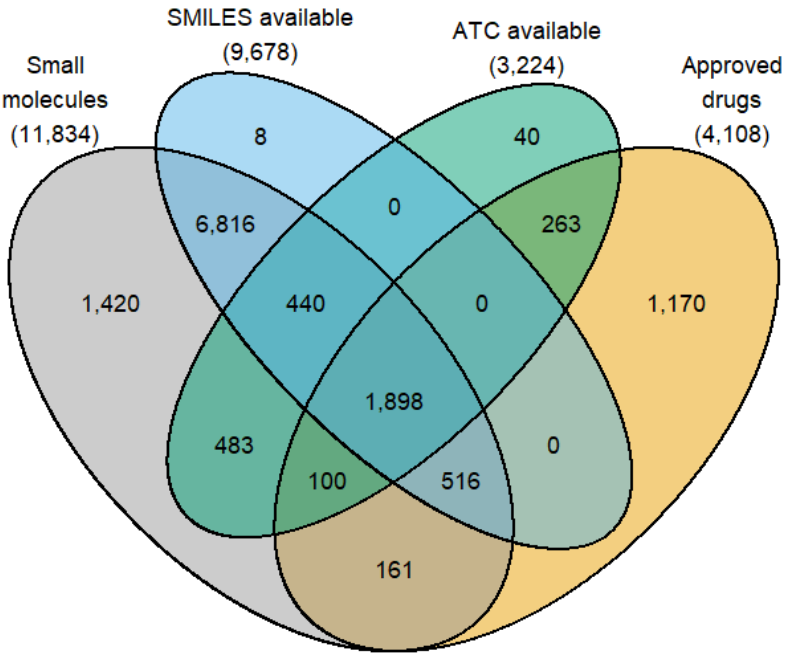


Fig. 4.2: A Venn diagram of the DrugBank filtering process. In total 1,898 drugs with unique drugbank id retrieved from DrugBank database were small molecule approved drugs and had both SMILES structure and ATC code available.

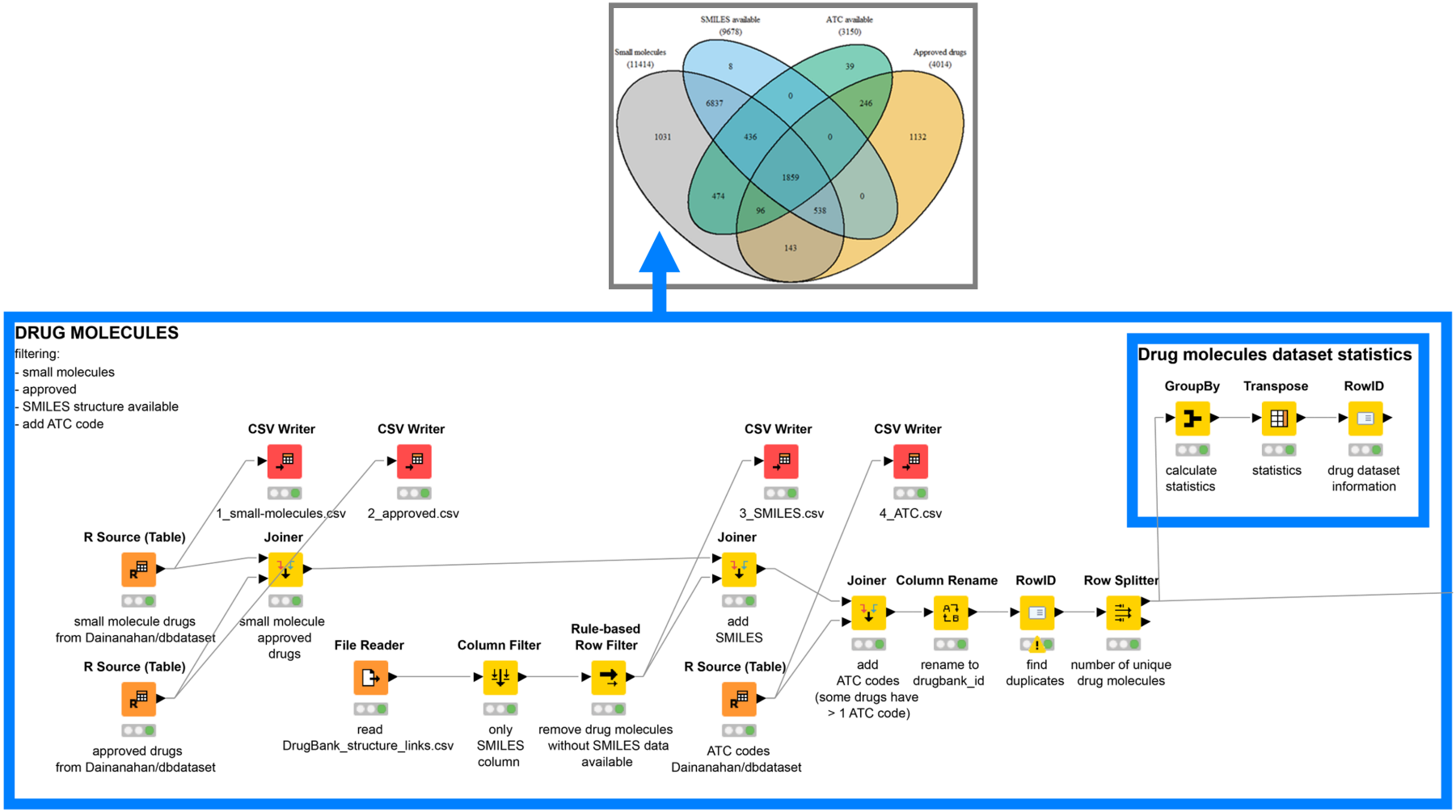


Fig. 4.3: The workflow for drug molecules retrieval and filtering

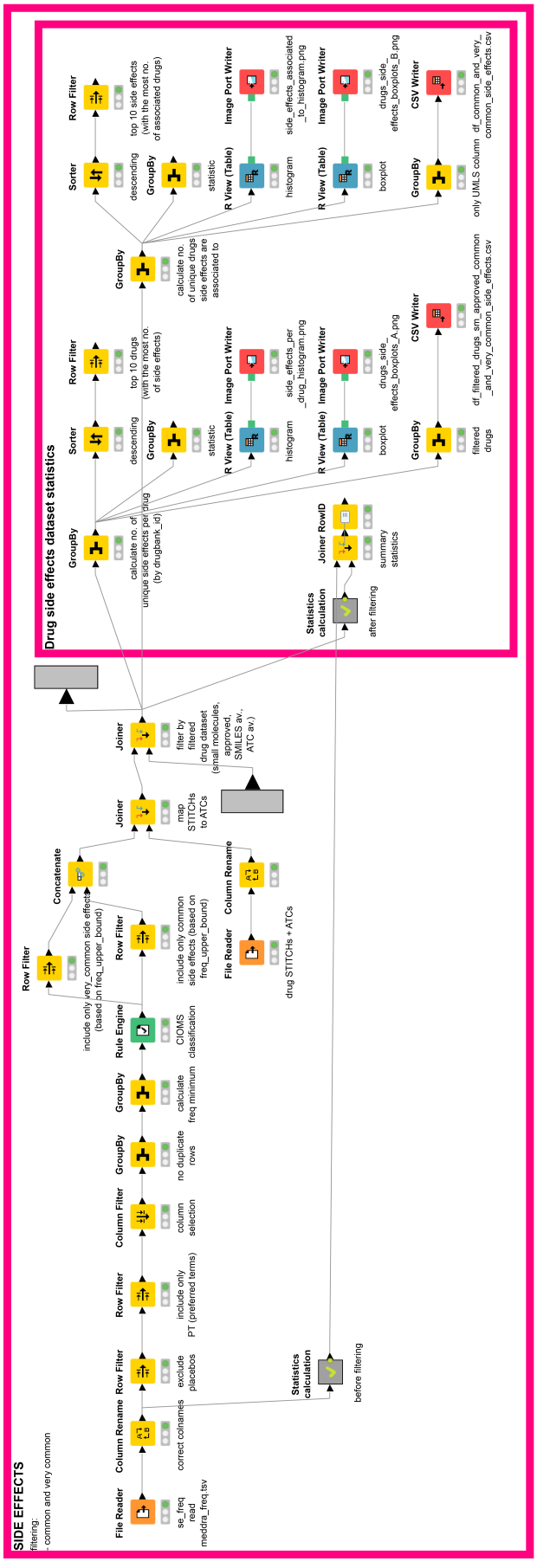


Fig. 4.4: The workflow for side effects retrieval and filtering

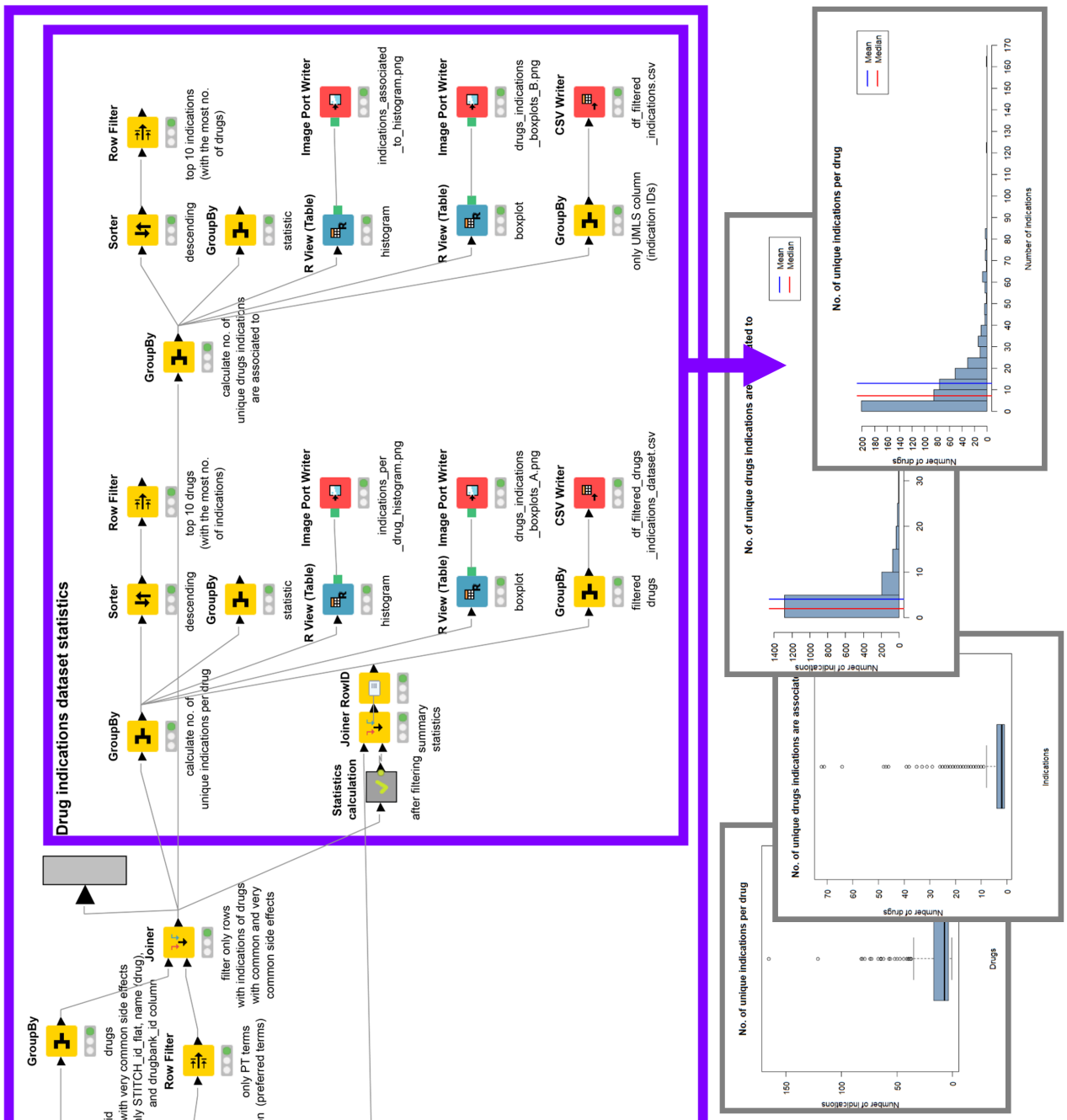


Fig. 4.5: The workflow for indications retrieval and filtering

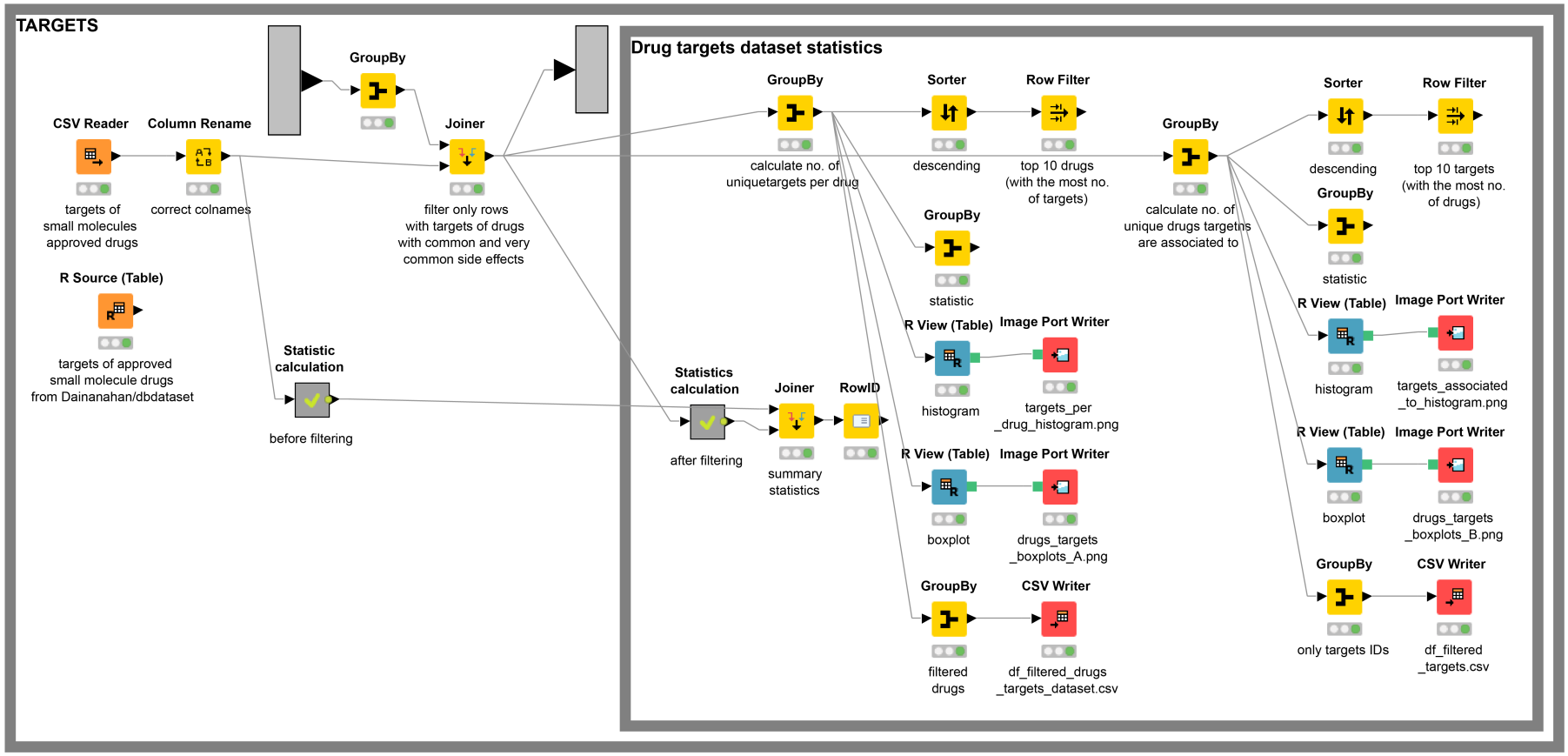


Fig. 4.6: The workflow for targets retrieval and filtering

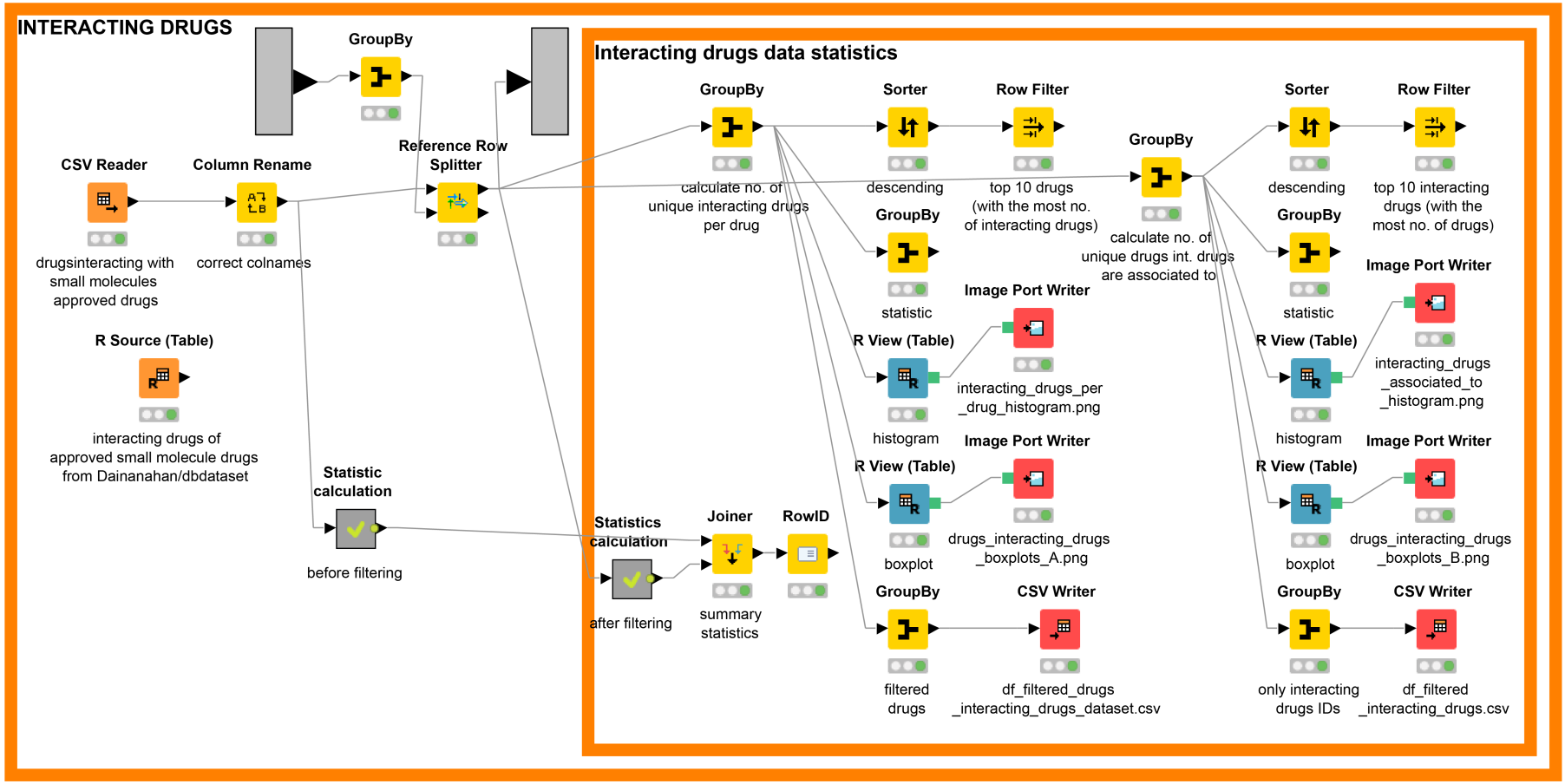


Fig. 4.7: The workflow for interacting drugs retrieval and filtering

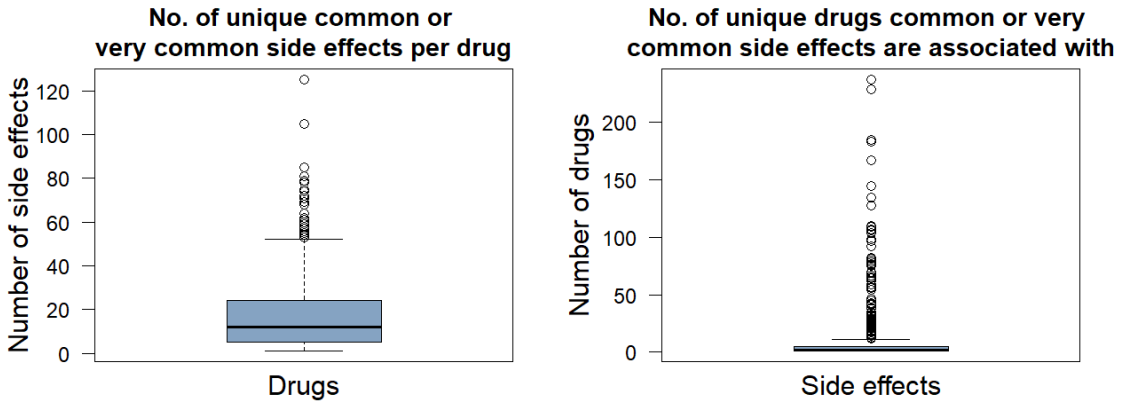


Fig. 4.8: The data distribution in the filtered side effect dataset

4.2.2 Drug side effects dataset after filtering

The size of the dataset was reduced only to examined drugs with side effects classified as 'common' and 'very common' based on the CIOMS classification (Tab. 2.3). As some DrugBankIDs can be associated with the same STITCH ID, we used DrugBankIDs as drug identifiers. Number of 511 drugs (drugbank_id terms) with at least 1 side effect remained in the dataset after filtering and 1,164 unique side effects were associated with those drugs, there were 9,114 drug–side effect associations in total (Tab. 4.3). The distribution of the number of unique side effects per examined drug and of the number of unique examined drugs with common or very common side effects are associated with is shown as box-and-whisker-plots (Fig. 4.8) and histograms (Fig. 4.9 and 4.10). The median of side effects per drug was 12 and the median of drugs each side effect is associated with was 2 (Tab. 4.4 and 4.6). Decitabine (DB01262) was labeled as a drug with the highest number of common or very common side effects among all drugs in the filtered dataset (Tab. 4.5). This medication associated with 125 side effects is indicated for the treatment of myelodysplastic syndromes (`https://go.drugbank.com/drugs/DB01262`). Headache (C0018681) was identified as the most prevalent side effect in the dataset (Tab. 4.7) with total 237 out of 511 drugs with this side effect.

Tab. 4.3: The drug side effect dataset information before and after filtering

Characteristics	**Before filtering**	**After filtering**
Number of drugs (unique STITCH_id_flat terms)	968	498
Number of drugs (unique drugbank_id terms)	NA	511
Number of side effects (unique UMLS_id terms)	3,775	1,164
Number of drug-SE associations (unique drug-SE_pairs)	55,932	9,114

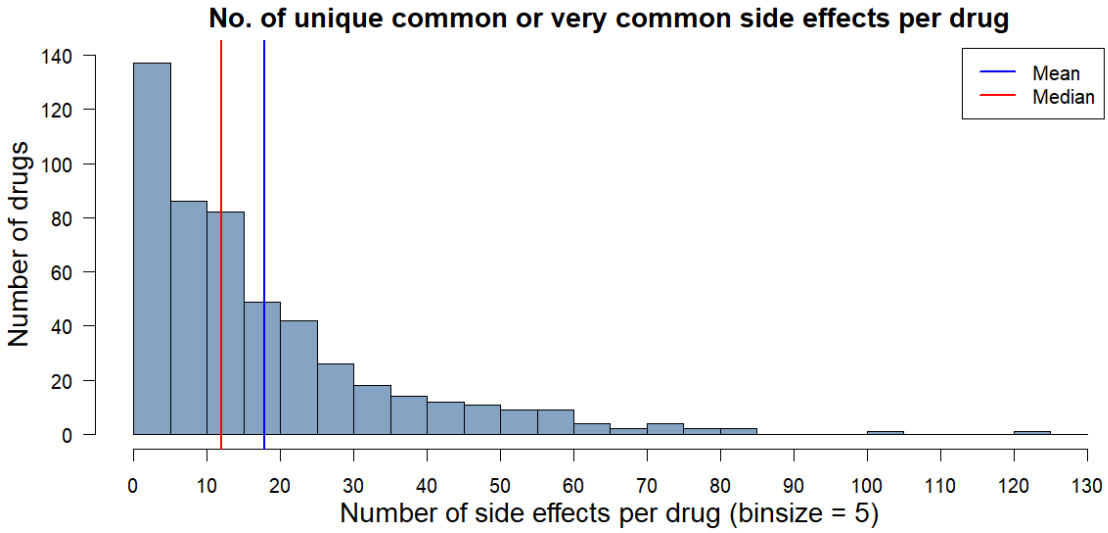


Fig. 4.9: The distribution of the number of unique side effects observed for each drug. The distribution is skewed right.

Tab. 4.4: The summary statistics of side effect distribution in the filtered dataset

Value	**No. of unique common and very common side effects per drug**
min	1
max	125
mean	17.84
median	12

Tab. 4.5: The top 10 drugs with the highest number of side effects

DrugBank ID (drug compound ID)	**Drug name**	**No. of common or very common side effects**
DB01262	Decitabine	125
DB01242	Clomipramine	105
DB00261	Anagrelide	85
DB00262	Carmustine	81
DB01224	Quetiapine	79
DB01202	Levetiracetam	78
DB00928	Azacitidine	75
DB00982	Isotretinoin	74
DB01610	Valganciclovir	72
DB01080	Vigabatrin	71

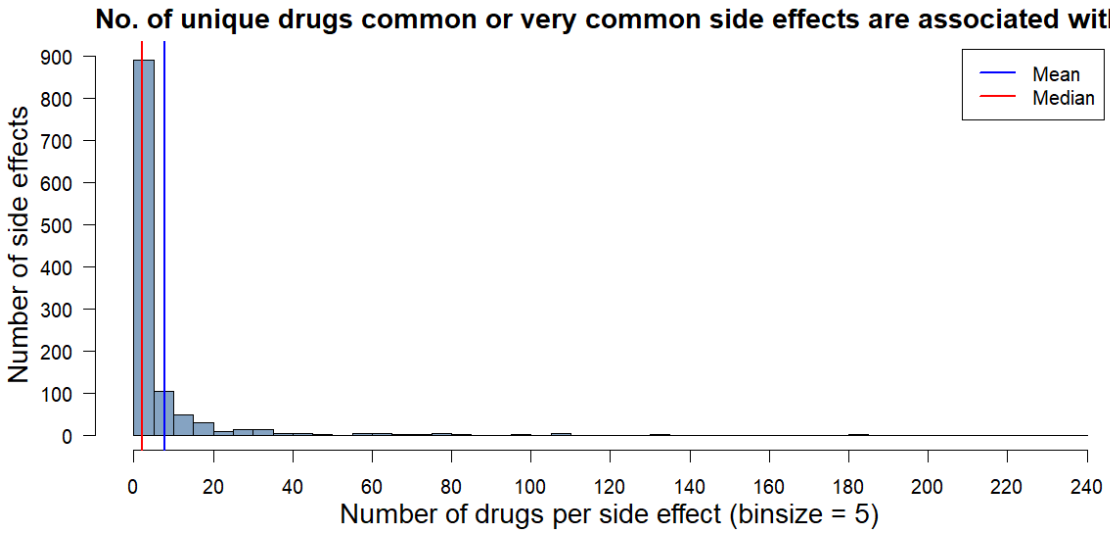


Fig. 4.10: The distribution of the number of unique drugs observed for each side effect. The distribution is skewed right.

Tab. 4.6: The summary statistics of drugs side effects are associated with

Value	**No. of unique drugs common and very common side effects are associated with**
min	1
max	237
mean	7.83
median	2

Tab. 4.7: The top 10 side effects associated with the highest number of drugs

UMLS (side effect ID)	**MedDRA_info_3 (side effect name(s))**	**No. of drugs with this side effect**
C0018681	Headache	237
C0027497	Nausea	229
C0012833	Dizziness	185
C0011991	Diarrhoea	183
C0042963	Vomiting	167
C0000737	Abdominal / Gastrointestinal pain	145
C0013395	Dyspepsia	135
C0015672	Asthenia, Fatigue	135
C0009806	Constipation	128
C0030193	Pain	110

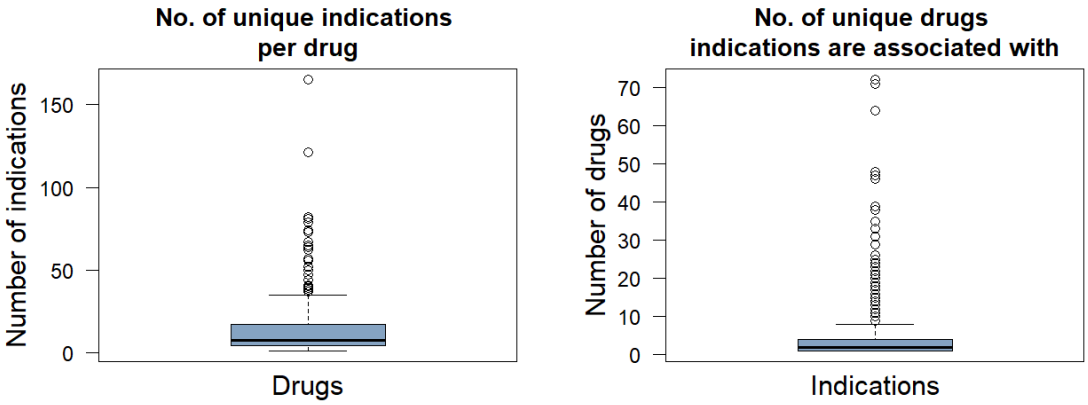


Fig. 4.11: The data distribution in the filtered indication dataset

4.2.3 Drug indications dataset after filtering

As indicated in the table below, 504 examined drugs (drugbank_id terms) with at least 1 indication remained in the filtered meddra_all_indications dataset and 1,618 unique indications were associated with these drugs, there were 6,696 drug–indication associations in total (Tab. 4.8). The distribution of the number of unique indications per examined drug and of the number of unique examined drugs indications are associated with is shown as box-and-whisker plots (Fig. 4.11) and histograms (Fig. 4.12 and 4.13). The median of the numbers of indications of each drug in the dataset was 7.5 and the median of the numbers of drugs each side effect is associated with was 2 (Tab. 4.9 and 4.11). Bethametasone (DB00443) was labeled as a drug with the highest number of indications among all drugs in the filtered dataset (Tab. 4.10). This medication is a systemic corticosteroid and is associated with 165 indications (`https://go.drugbank.com/drugs/DB00443`). Renal failure (C0035078) was identified as the most prevalent indication in the dataset (Tab. 4.12). Total number of drugs with this indication was 72 out of 504.

Tab. 4.8: The drug indications dataset information before and after filtering

Characteristics	**Before filtering**	**After filtering**
Number of drugs (unique STITCH_id_flat terms)	1,437	491
Number of drugs (unique drugbank_id terms)	NA	504
Number of indications (unique UMLS_id_MedDRA_indication terms)	3,046	1,618
Number of drug-indication associations (unique drug-indication_pairs)	17,879	6,696

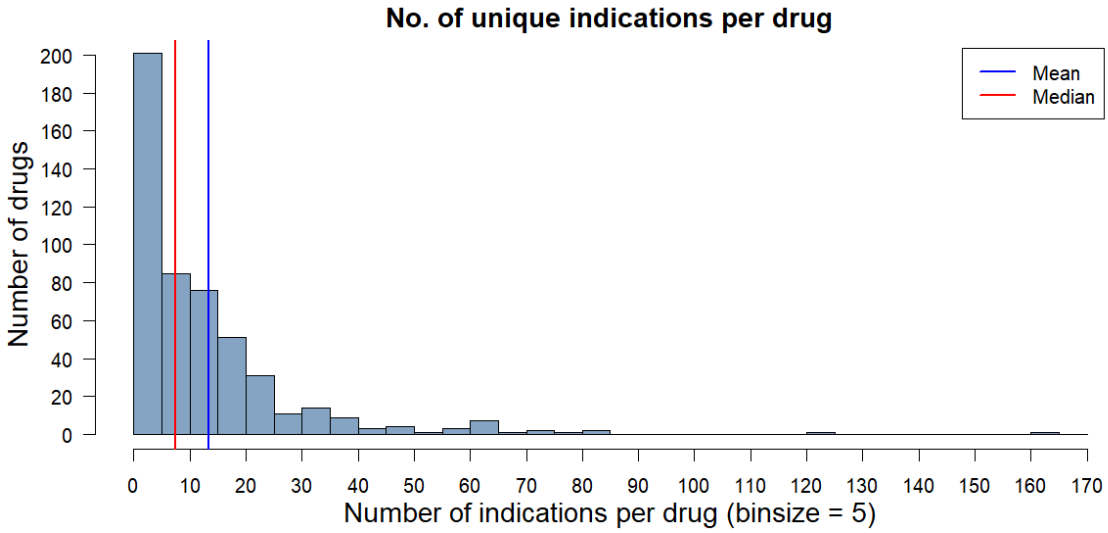


Fig. 4.12: The distribution of the number of unique indications observed for each drug. The distribution is skewed right.

Tab. 4.9: The summary statistics of indications distribution in the filtered dataset

Value	**No. of unique indications per drug**
min	1
max	165
mean	13.29
median	7.5

Tab. 4.10: The top 10 drugs in the filtered with the highest number of indications

DrugBank ID (drug compound ID)	**Drug name**	**No. of indications**
DB00443	Betamethasone	165
DB00741	Hydrocortisone	121
DB00254	Doxycycline	82
DB00715	Paroxetine	81
DB00620	Triamcinolone	79
DB01017	Minocycline	74
DB01104	Sertraline	73
DB00404	Alprazolam	67
DB00424	Hyoscyamine	65
DB00572	Atropine	65

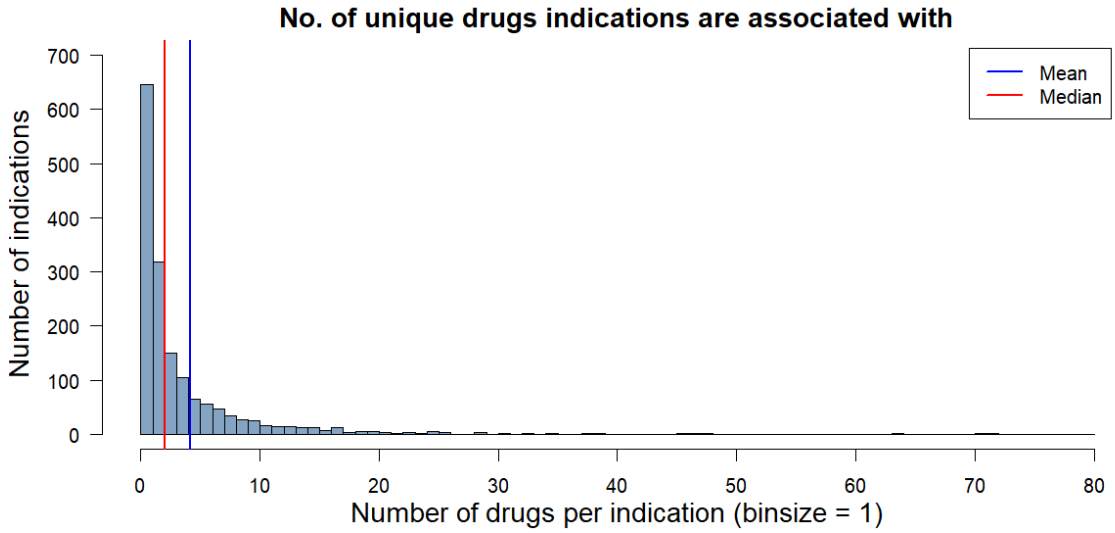


Fig. 4.13: The distribution of the number of unique drugs observed for each indication. The distribution is skewed right.

Tab. 4.11: The statistics of drugs indications are associated with

Value	**No. of unique drugs indications are associated with**
min	1
max	72
mean	4.14
median	2

Tab. 4.12: Top 10 indications associated to the highest number of drugs

UMLS id MedDRA indication (indication ID)	**concept_name_MedDRA (indication name)**	**No. of drugs with this indication**
C0035078	Renal failure	72
C0009450	Infection	71
C1565489	Renal impairment	71
C0006826	Neoplasm malignant	64
C0027651	Neoplasm	48
C0023895	Liver disorder	47
C0030193	Pain	47
C0011849	Diabetes mellitus	46
C0020538	Hypertension	46
C0036572	Convulsion	39

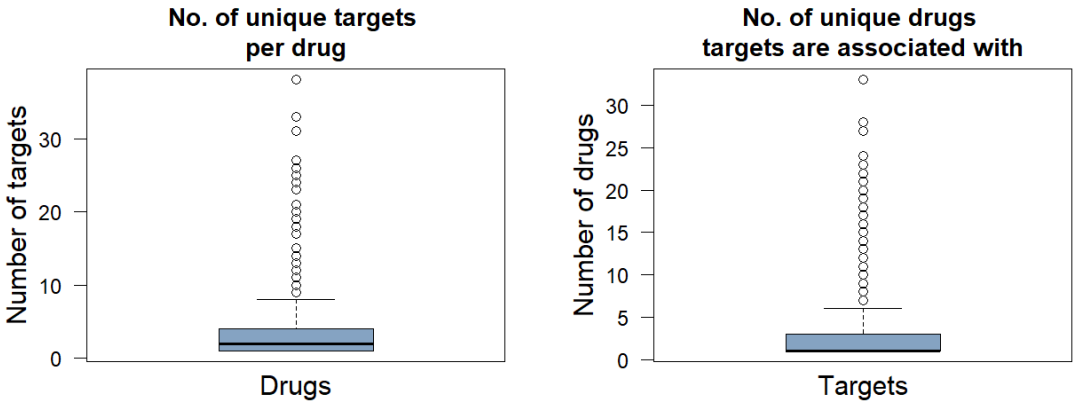


Fig. 4.14: The data distribution in filtered target dataset

4.2.4 Drug targets dataset after filtering

As indicated in the table below, 482 examined drugs (drugbank_id terms) with at least 1 target remained in the filtered target dataset and 646 unique targets were associated with these drugs, there were 1,867 drug–target associations in total (Tab. 4.13). The distribution of the number of unique targets per examined drug and of the number of unique examined drugs targets are associated with is shown as box-and-whisker plots and histograms (Fig. 4.15 and 4.16). The median of the numbers of targets of each drug in the dataset was 2 and the median of the numbers of drugs each target is associated with was 1 (Tab. 4.14 and 4.16). Aripiprazole (DB01238) was labeled as a drug with the highest number of targets among all drugs in the filtered dataset (Tab. 4.15). This medication is an atypical antipsychotic and is associated with 38 targets (`https://go.drugbank.com/drugs/DB01238`). DNA (BE0004796) was identified as the most prevalent target in the dataset (Tab. 4.17). Total number of drugs with this target was 33 out of 482.

Tab. 4.13: Drug targets dataset information before and after filtering

Characteristics	**Before filtering**	**After filtering**
Number of drugs (unique drugbank_id terms)	1,954	482
Number of targets (unique target_id terms)	2,498	646
Number of drug-target associations (unique drug-target_pairs)	8,278	1,867

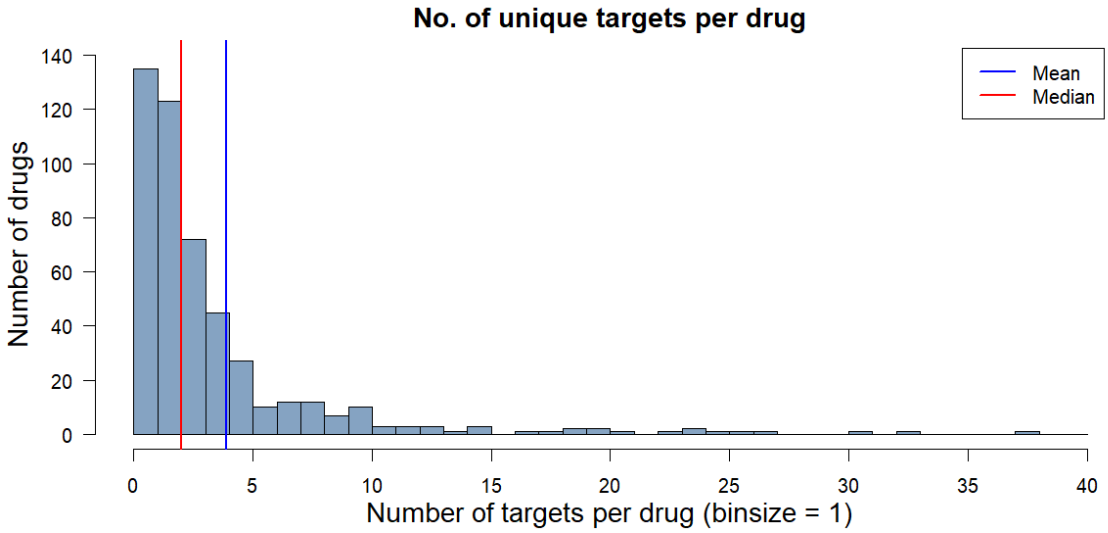


Fig. 4.15: The distribution of the number of unique targets observed for each drug. The distribution is skewed right.

Tab. 4.14: The summary statistics of the targets distribution in the filtered dataset

Value	**No. of unique indications per drug**
min	1
max	38
mean	3.87
median	2

Tab. 4.15: The top 10 drugs with the highest number of targets

DrugBank ID (drug compound ID)	**Drug name**	**No. of targets**
DB01238	Aripiprazole	38
DB00408	Loxapine	33
DB00909	Zonisamide	31
DB00363	Clozapine	27
DB00246	Ziprasidone	26
DB00543	Amoxapine	25
DB00458	Imipramine	24
DB01224	Quetiapine	24
DB01254	Dasatinib	23
DB00248	Cabergoline	21

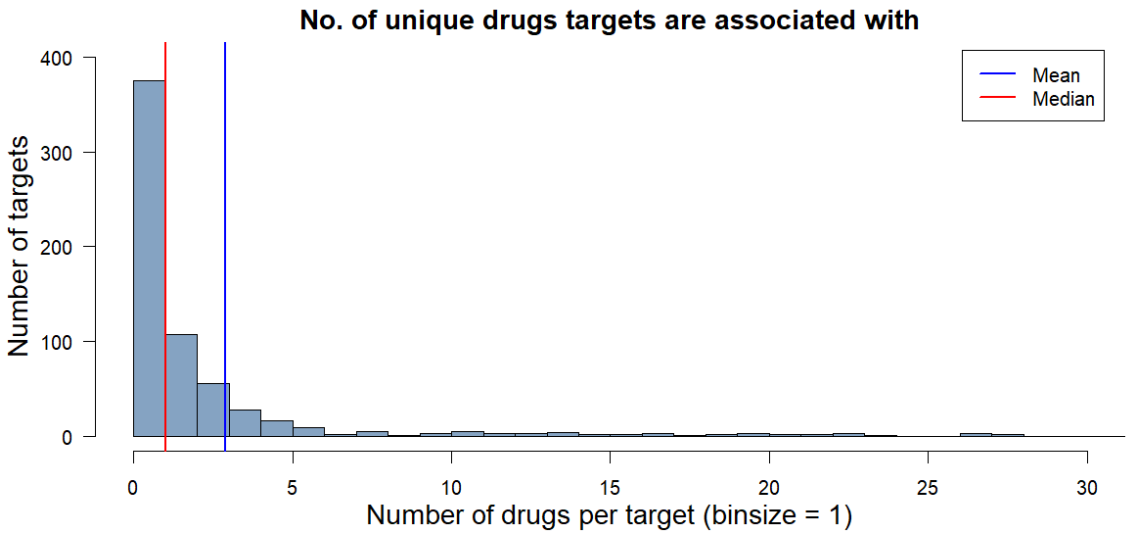


Fig. 4.16: The distribution of the number of unique drugs observed for each target. The distribution is skewed right.

Tab. 4.16: The statistics of drugs targets are associated with

Value	**No. of unique drugs indications are associated with**
min	1
max	33
mean	2.89
median	1

Tab. 4.17: The top 10 targets associated with the highest number of drugs

Target (target ID)	**Target name**	**No. of drugs with this target**
BE0004796	DNA	33
BE0000442	Histamine H1 receptor	28
BE0000756	Dopamine D2 receptor	28
BE0000291	5-hydroxytryptamine receptor 1A	27
BE0000451	5-hydroxytryptamine receptor 2A	27
BE0000501	Alpha-1A adrenergic receptor	27
BE0000092	Muscarinic acetylcholine receptor M1	24
BE0000560	Muscarinic acetylcholine receptor M2	23
BE0000045	Muscarinic acetylcholine receptor M3	23
BE0000533	5-hydroxytryptamine receptor 2C	23

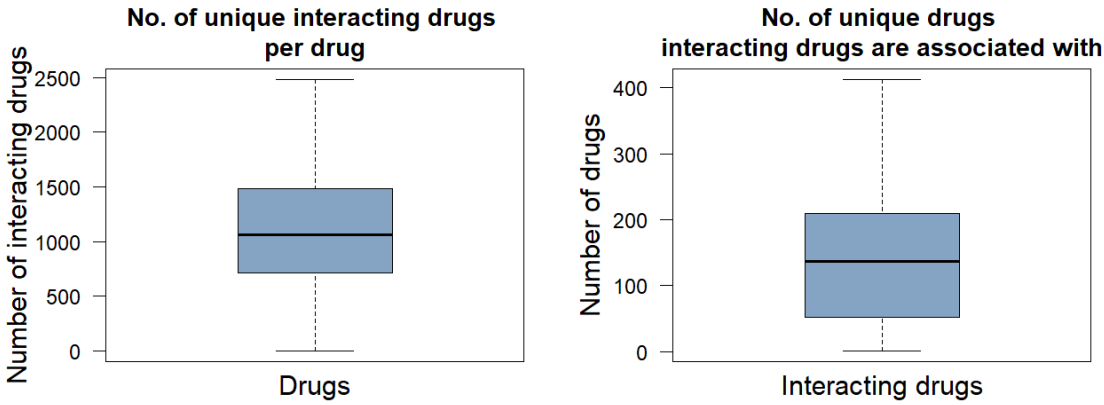


Fig. 4.17: The data distribution in the filtered interacting drugs dataset

4.2.5 Interacting drugs dataset after filtering

As indicated in the table below, 469 examined drugs (drugbank_id terms) with at least 1 interacting drug remained in the filtered interacting drugs dataset and 3,548 unique interacting drugs were associated with those drugs, there were 494,185 drug–drug associations in total (Tab. 4.18). The distribution of the number of unique interacting drugs per examined drug and of the number of unique examined drugs interacting drugs are associated with is shown as box-and-whisker plots (Fig. 4.17) and histograms (Fig. 4.18 and 4.19). The median of the numbers of interacting drugs of each examined drug in the dataset was 1,062 and the median of the numbers of examined drugs each interacting drug is associated with was 138 (Tab. 4.19 and 4.21). Quinidine (DB00908) was labeled as a drug with the highest number of interacting drugs among all drugs in the filtered dataset (Tab. 4.20). This medication is indicated for the treatment of ventricular pre-excitation and cardiac dysrhythmia (`https://go.drugbank.com/drugs/DB00908`) and is associated with 2,477 interacting drugs. Quinidine was also identified as the most interacting drug in the dataset (Tab. 4.22). Total number of drugs associated with this interacting drug was 412 out of 469.

Tab. 4.18: The drug interacting drugs dataset information before and after filtering

Characteristics	**Before filtering**	**After filtering**
Number of drugs (unique drugbank_id terms)	2,090	469
Number of interacting drugs (unique drugbank_id_int terms)	4,063	3,548
Number of drug-drug associations (unique drug-drug_pairs)	1,722,202	494,185

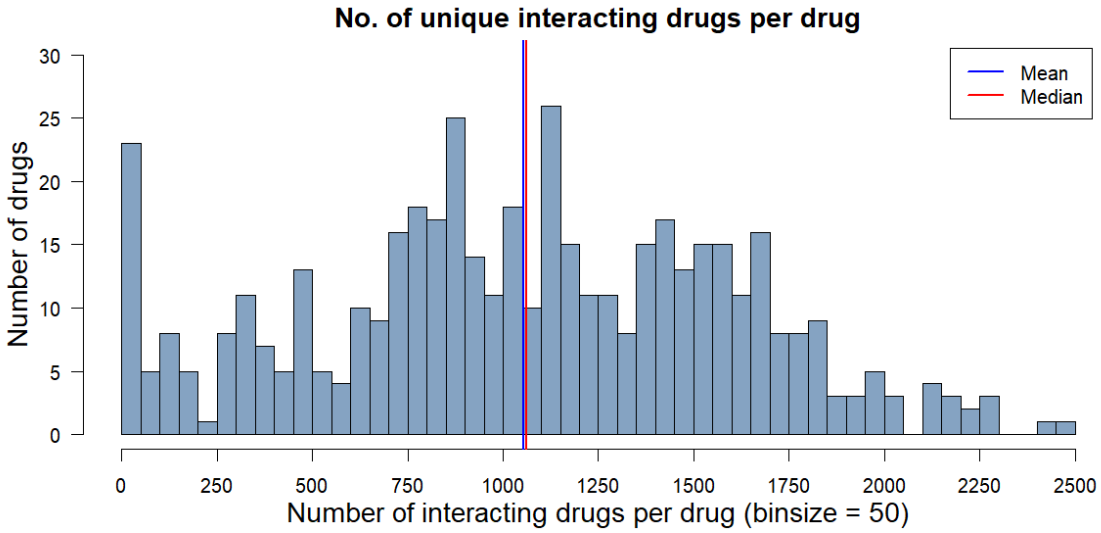


Fig. 4.18: The distribution of the number of unique interacting drugs observed for each drug. The distribution is random and has several peaks.

Tab. 4.19: The summary statistics of the interacting drugs distribution

Value	**No. of unique interacting drugs per drug**
min	1
max	2,477
mean	1,053.7
median	1,062

Tab. 4.20: The top 10 drugs with the highest number of interacting drugs

DrugBank ID (drug compound ID)	**Drug name**	**No. of interacting drugs**
DB00908	Quinidine	2,477
DB00363	Clozapine	2,437
DB01142	Doxepin	2,273
DB00564	Carbamazepine	2,270
DB00458	Imipramine	2,260
DB00476	Duloxetine	2,215
DB00909	Zonisamide	2,210
DB00280	Disopyramide	2,178
DB01224	Quetiapine	2,170
DB00091	Cyclosporine	2,167

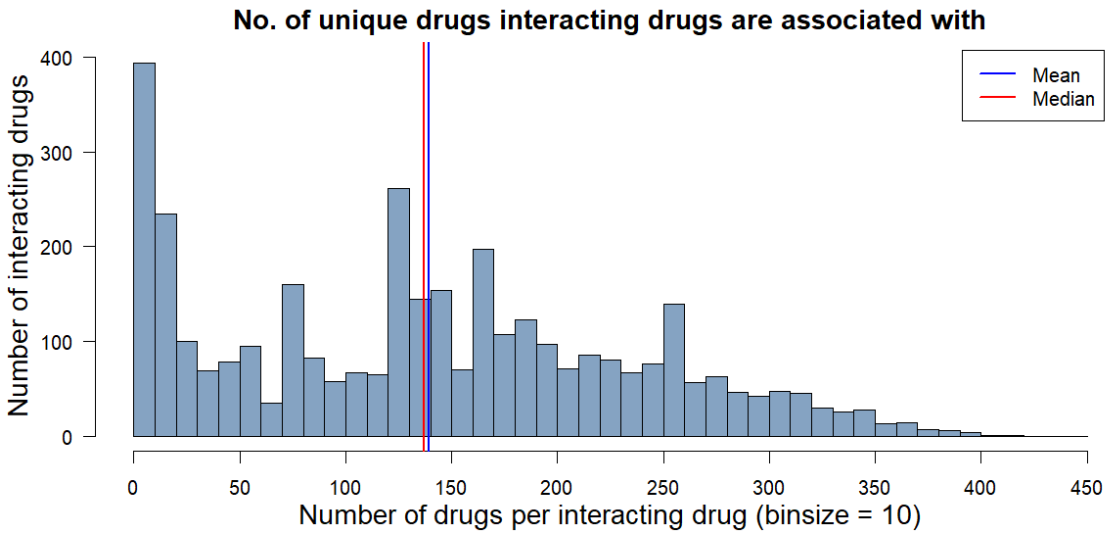


Fig. 4.19: The distribution of the number of unique drugs observed for each interacting drug. The distribution is skewed right.

Tab. 4.21: The statistics of the drugs interacting drugs are associated with

Value	**No. of unique drugs targets are associated with**
min	1
max	412
mean	139.29
median	137

Tab. 4.22: Top 10 interacting drugs associated to the highest number of drugs

Interacting drug (interacting drug ID)	**Interacting drug name**	**No. of drugs with this interacting drug**
DB00908	Quinidine	412
DB00363	Clozapine	409
DB00091	Cyclosporine	400
DB00477	Chlorpromazine	400
DB00502	Haloperidol	396
DB00398	Sorafenib	392
DB01151	Desipramine	390
DB00321	Amitriptyline	389
DB00564	Carbamazepine	384
DB00333	Methadone	383

4.3 Similarity metrics calculation

The following workflows demonstrate a similarity metrics calculation between each drug pair in the filtered datasets. The first workflow is intended for a structure similarity calculation (Fig. 4.20). The workflow begins with a structural drug data transformation which is performed within the 'Preprocessing' metanode on the filtered drug structure database. In the metanode strings representing the examined drug structures are converted into KNIME SMILES strings via 'Molecule Type Cast' KNIME node which casts a string as a chemical type. Then, we converted the structures into the RDKit internal representation using an 'RDKit From Molecule' KNIME node.

Next, multiple fingerprints were generated for all drugs of loaded dataset within 'Fingerprints' metanode in which multiple 'RDKit fingerprint' KNIME nodes were used to calculate different types of molecule structure fingerprints. In total, 8 unique fingerprints were generated for all drug molecules: the Morgan fingerprint, the Feat-Morgan fingerprint, the AtomPair fingerprint, the Torsion fingerprint, the RD-Kit fingerprint, the Avalon fingerprint, the Layered fingerprint and the MACCS fingerprint.

The fingerprint generation was followed by a fingerprint similarity calculation loop. In each iteration, one query drug molecule was compared to remaining molecules in the dataset. The 'Fingerprint Similarity' metanode computes the similarity of molecules by calculating the Tanimoto similarity coefficient based on the CDK toolkit for each fingerprint type. The resulting similarity values were exported as .csv files using 'CSV writer' KNIME node.

In the remaining four similarity calculation metanodes, the Jaccard indexes calculation was performed for side effect similarity, targets similarity, indications similarity, and interacting drugs similarity. Figure 4.21 illustrates an example workflow which allows the Jaccard similarity calculation for shared targets. At first, sets of targets were converted into bit vectors fingerprints (target fingerprints). The fingerprint generation was followed by a Jaccard similarity calculation nested loop. In each iteration, one query drug target fingerprint was compared to the remaining drug target fingerprints in the dataset and the similarity metrics was calculated by a set of necessary operations. The other Jaccard similarity indexes were calculated in a similar way.

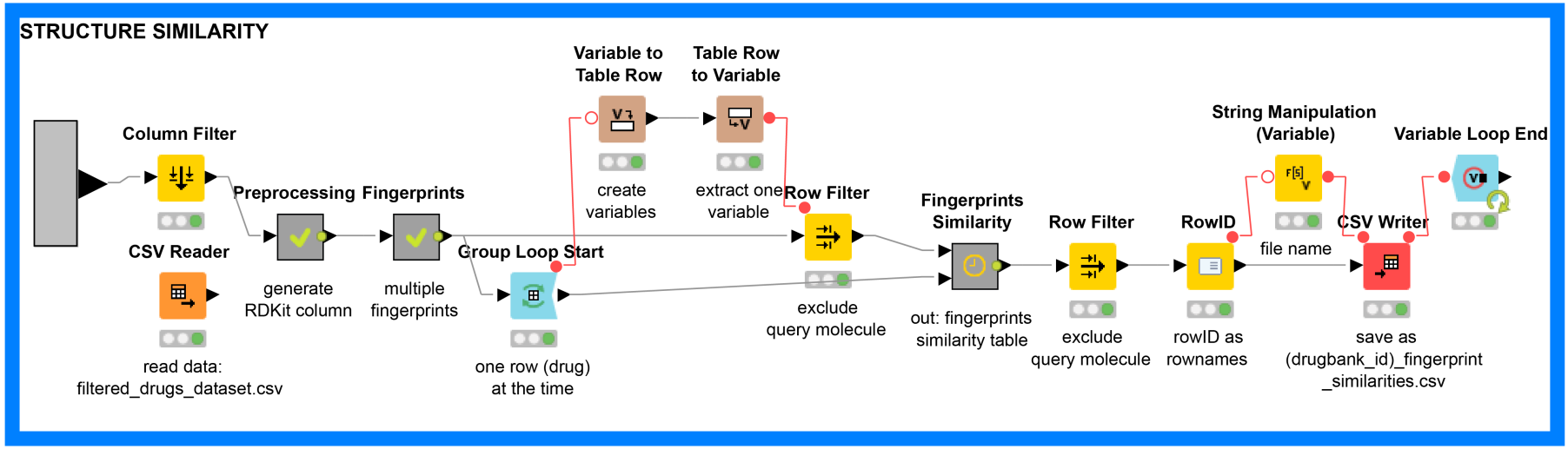


Fig. 4.20: Workflow for Tanimoto structure similarity coefficients calculation. In total, 8 unique fingerprints were generated for all drug molecules within the 'Fingerprints' metanode. Fingerprint similarity calculation has been performed in the 'Fingerprint Similarity' metanode for each fingerprint type.

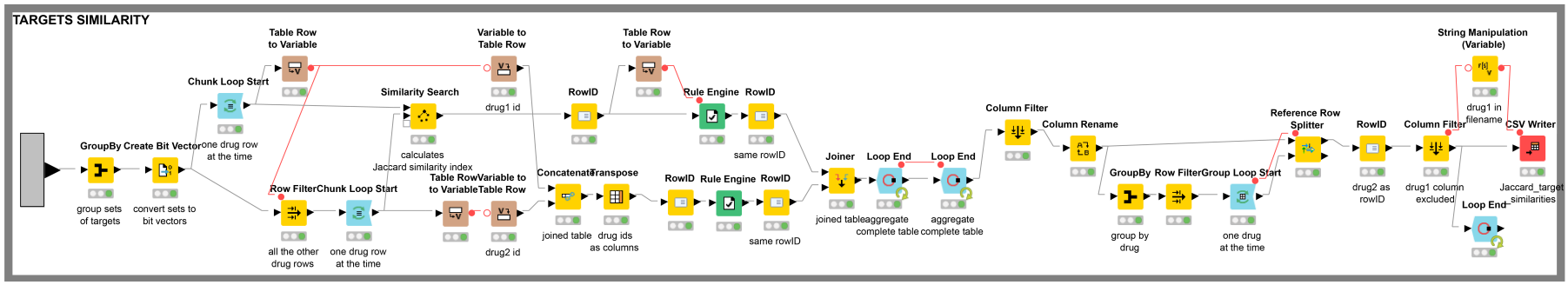


Fig. 4.21: An example workflow for the Jaccard similarity indexes calculation - calculation of the shared targets similarity. Sets of targets were converted into bit vectors fingerprints and in a following nested loop, one query drug target fingerprint was compared to the remaining drug target fingerprints of the dataset and the Jaccard similarity metrics was calculated in each iteration.

4.4 Datasets for analysis preparation and exploration

The second part of the set of the workflows deals with datasets for analysis preparation and exploration and a selection datasets creation (Fig. 4.22).

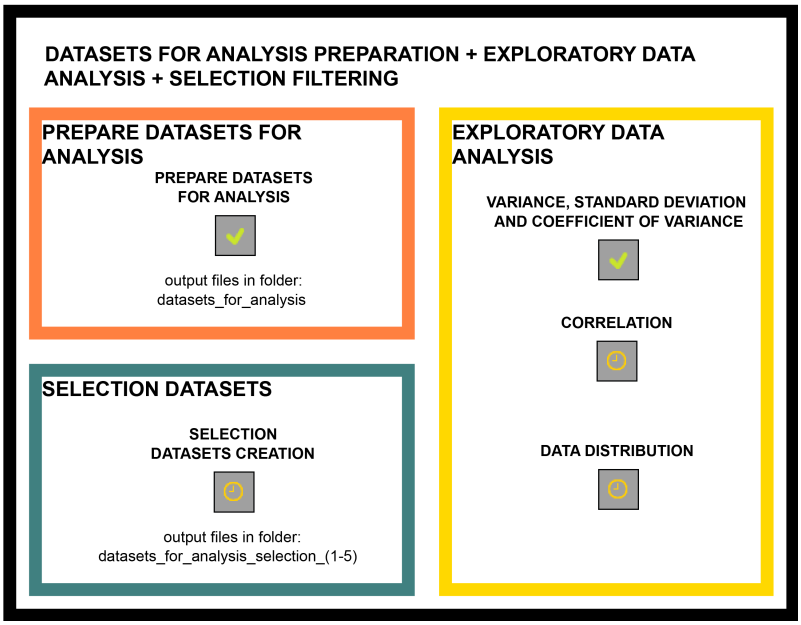


Fig. 4.22: The workflow overview - part II: a datasets for analysis preparation, an exploratory data analysis and a selection datasets creation.

All the similarity metrics prepared in the first part were combined together using the workflow depicted below (Fig. 4.23). The similarity coefficients data were joined with an association with a given side effect data. The dataframes of the drugs merged with the similarity coefficients and a labeled association with a given side effect (positive or negative) result from this workflow and compile datasets for machine-learning analysis. Besides that, the datatables corresponding to the query drug and side effect association with tested in a later phase were extracted in each iteration.

Several rows of prepared example dataset for analysis can be found in the table below (Tab. 4.23). The rows of dataset for analysis represent the drug molecules. The columns represent structure fingerprint similarities and the Jaccard similarity indexes between molecule in a given row and the query molecule. The last column represent query side effect association. The values represent similarity of each drug molecule in the dataset to a query molecule. The columns represent the Tanimoto structure similarity coefficients for the different fingerprint types (Morgan, Feat-Morgan, AtomPair, Torsion, RD-Kit, Avalon, Layered, MACCS) and the Jaccard similarity index for the shared side effects, the shared indications, the shared targets or the shared drugs. As a consequence, a total number of 4,690 datasets for analysis were constructed.

Tab. 4.23: An example of prepared dataset for the analysis (combination for drug DB00006 and side effect C0000737). Columns represent similarity coefficients for different fingerprint types.

DrugBank ID	RDKit Tanimoto	Feat Morgan Tanimoto	Atom Pair Tanimoto	Torsion Tanimoto	Avalon Tanimoto	Layered Tanimoto	MACCS Tanimoto	Pattern Tanimoto	Jaccard index side effects	Jaccard index indications	Jaccard index targets	Jaccard index interacting drugs	Side effect association
DB00014	0.74	0.32	0.93	0.48	0.45	0.74	0.80	0.81	0.06	0.00	0.00	0.17	negative
DB00080	0.73	0.35	0.97	0.41	0.46	0.70	0.72	0.85	0.00	0.00	0.00	0.21	negative
DB00091	0.64	0.25	0.78	0.28	0.51	0.63	0.64	0.77	0.00	0.00	0.00	0.2	negative
DB00104	0.69	0.29	0.76	0.37	0.36	0.66	0.62	0.68	0.00	0.04	0.00	0.13	positive
DB00136	0.36	0.08	0.30	0.13	0.17	0.45	0.36	0.46	0.00	0.00	0.00	0.14	negative
DB00175	0.44	0.19	0.32	0.10	0.26	0.43	0.40	0.41	0.08	0.11	0.00	0.14	negative

4.4.1 Data exploration

Before proceeding to machine-learning, an exploratory data analysis was performed as the next step. The data distribution analysis involves a workflow which aims at features exploration in terms of statistical measures for potential dimensionality reduction as described in the methods section. The workflow computes the following measures which are important for feature characteristics: variance, standard deviation, coefficient of variation (Fig. 4.24). The resulting values can be visualized as box-and-whisker plots via integrated R scripts in 'R View (Table)' KNIME nodes. A feature correlation was calculated and visualized in an additional workflow (Fig. 4.25). The results of the exploratory data analysis are described in the following figures.

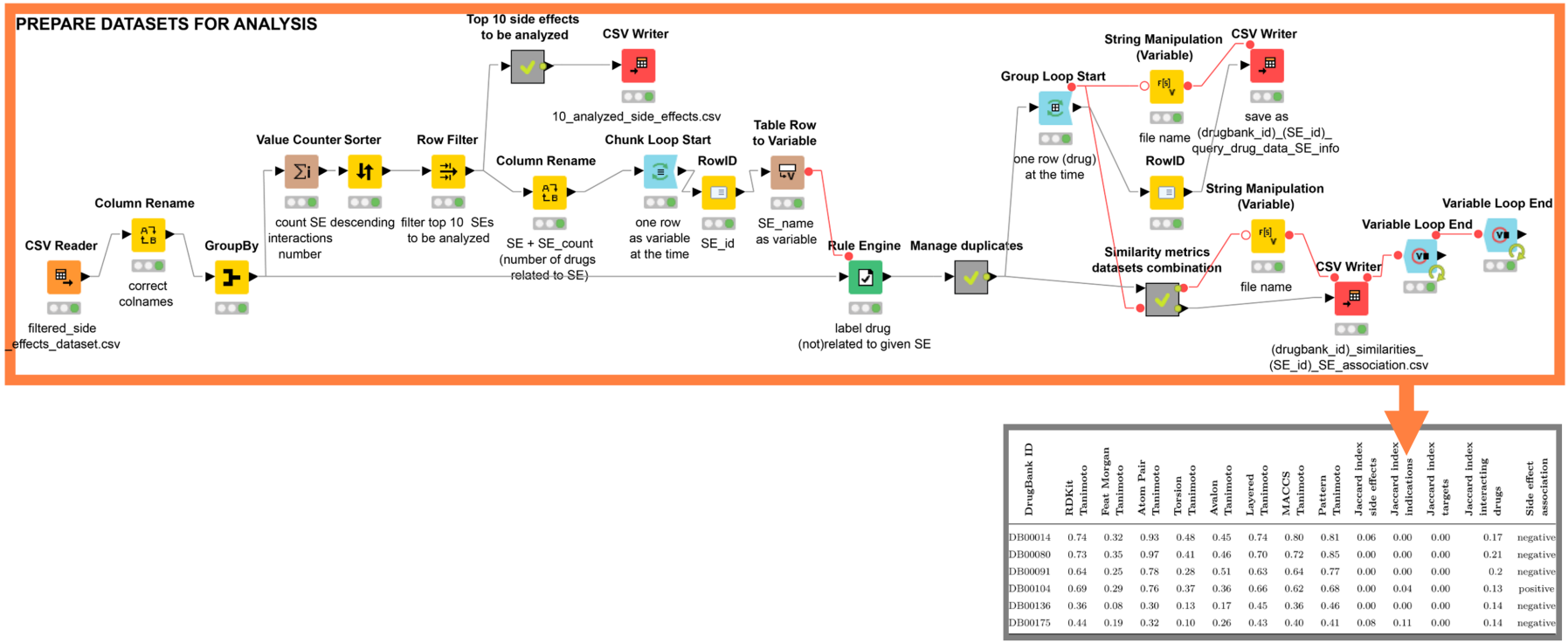
As the box-and-whisker plots below show, our variables are not very good in terms of variance (Fig. 4.26). The RDKit Tanimoto similarity and the Pattern Tanimoto similarity represent the best candidate features. From the Jaccard similarities, the interacting drugs Jaccard similarity index feature shows the best results regarding data variance.

As shown in the figure below (Fig. 4.27), again the RDKit Tanimoto similarity and the Pattern Tanimoto similarity features have a greater median of standard deviation than all the other features. Similarly, the median of the standard deviation of the interacting drugs similarity feature is highest and hence there is the most variation and this feature is supposed to have the best predictive power. The standard

deviation in most of the other features is close to 0 which indicates that the data tend to be close to the mean.

The correlation for a specific dataset can be visualized in a correlogram separately. The example relationship between the fingerprint similarity variables is shown in the appendix (Fig. B.1). The figure combining correlogram with the significance test is intended to investigate dependence between all similarities at the same time.

However, we focused on the feature correlation distribution visualization within all datasets. For resulting histograms see the appendix. The correlation coefficients show that some of the features are more positively correlated than others. The median values of the correlations range from 0.1 to 0.83. According to a hypothesis, one of the two very correlated columns can be removed without decreasing the amount of information for further analysis. Based on the resulting observations user can decide to avoid specific features if they are supposed not to have much predictive power. We decided to remove the Avalon and the Layered feature columns as their median values showed a strong correlation above 0.7 with the highest number of other features as indicated in the table in the appendix (Tab. B.1).



DrugBank ID	RDKit Tanimoto	Feat Morgan Tanimoto	Atom Pair Tanimoto	Torsion Tanimoto	Avalon Tanimoto	Layered Tanimoto	MACCS Tanimoto	Pattern Tanimoto	Jaccard index side effects	Jaccard index indications	Jaccard index targets	Jaccard index interacting drugs	Side effect association
DB00014	0.74	0.32	0.93	0.48	0.45	0.74	0.80	0.81	0.06	0.00	0.00	0.17	negative
DB00080	0.73	0.35	0.97	0.41	0.46	0.70	0.72	0.85	0.00	0.00	0.00	0.21	negative
DB00091	0.64	0.25	0.78	0.28	0.51	0.63	0.64	0.77	0.00	0.00	0.00	0.2	negative
DB00104	0.69	0.29	0.76	0.37	0.36	0.66	0.62	0.68	0.00	0.04	0.00	0.13	positive
DB00136	0.36	0.08	0.30	0.13	0.17	0.45	0.36	0.46	0.00	0.00	0.00	0.14	negative
DB00175	0.44	0.19	0.32	0.10	0.26	0.43	0.40	0.41	0.08	0.11	0.00	0.14	negative

Fig. 4.23: The workflow combining all similarity metrics and creating datasets for analysis including the side effect association column. Additionally, a list of top 10 side effects is filtered.

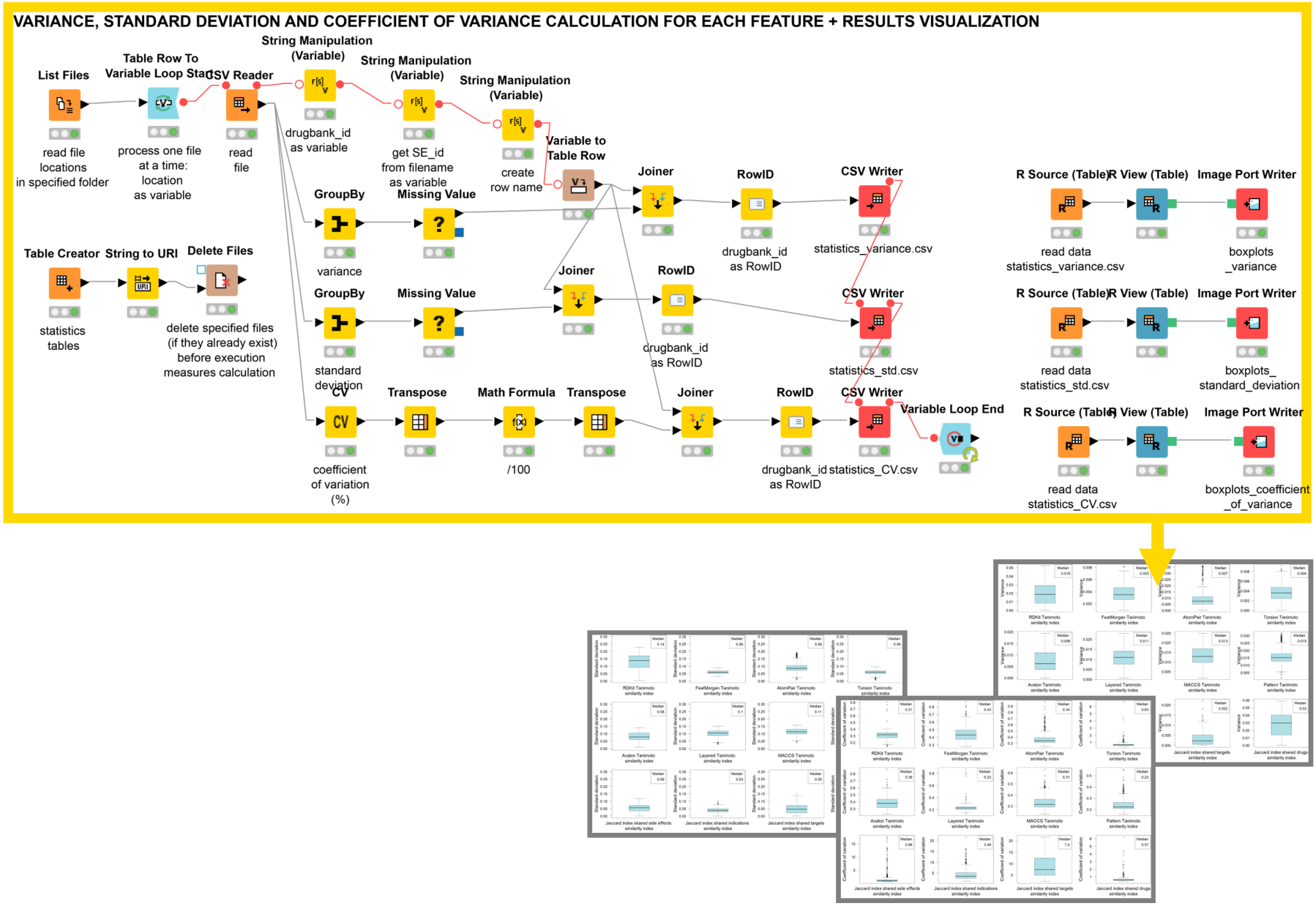


Fig. 4.24: The workflow to compute and visualize measures for the feature selection. The resulting value distribution is visualized as box-and-whisker plots.

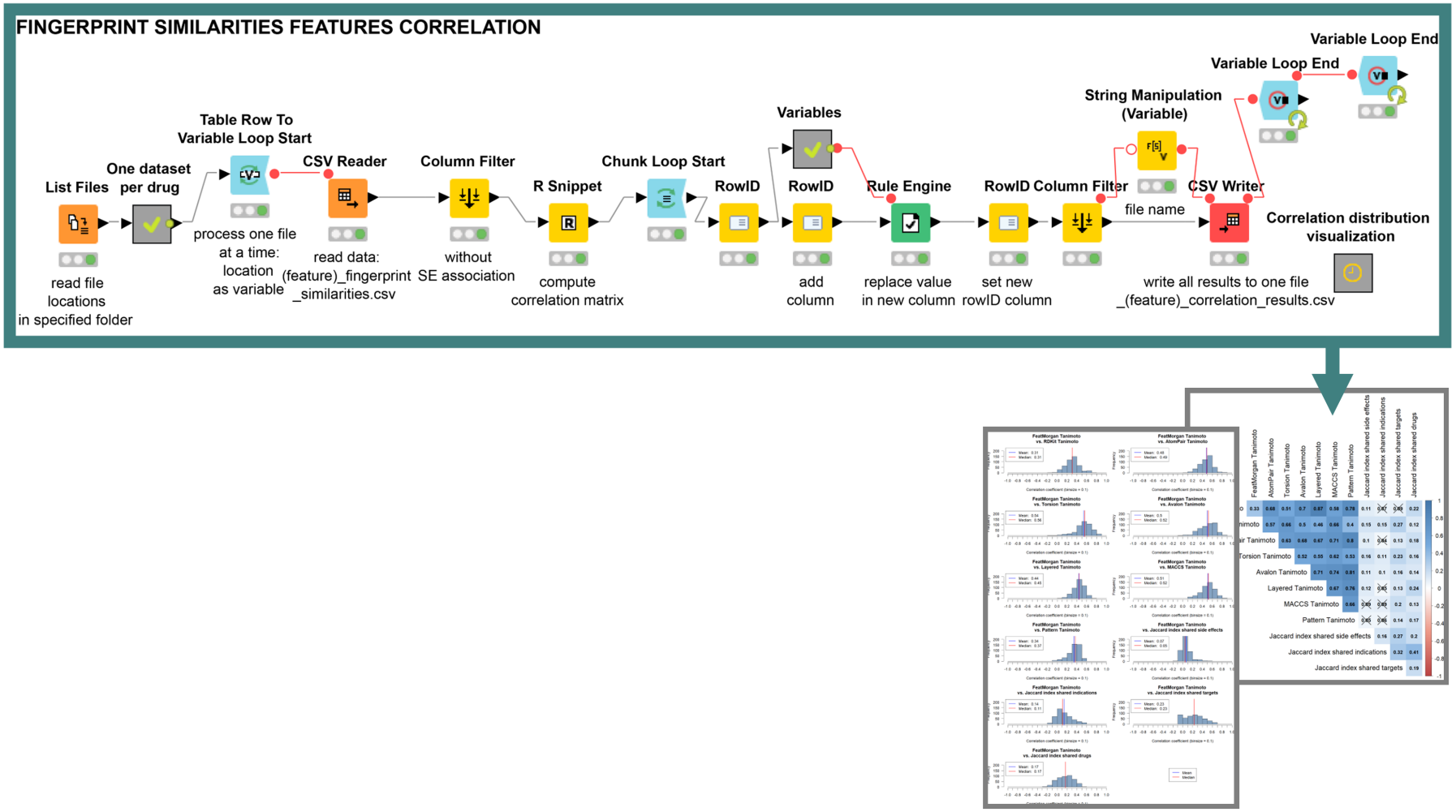


Fig. 4.25: The workflow to compute and visualize correlation between the similarity values. The results can be plotted as correlograms or histograms via integrated R scripts.

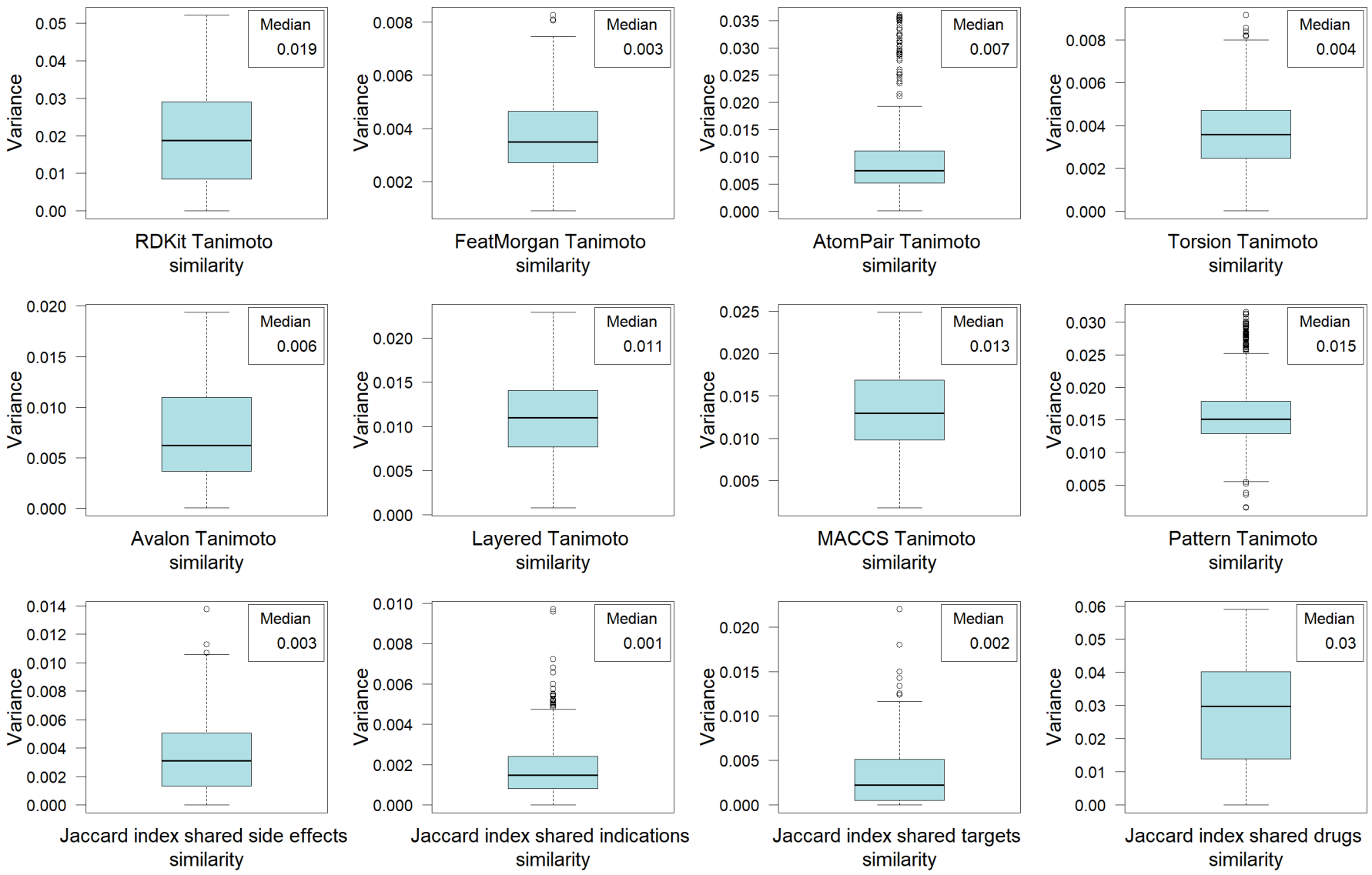


Fig. 4.26: The box-and-whisker plots showing the variance distribution for each feature

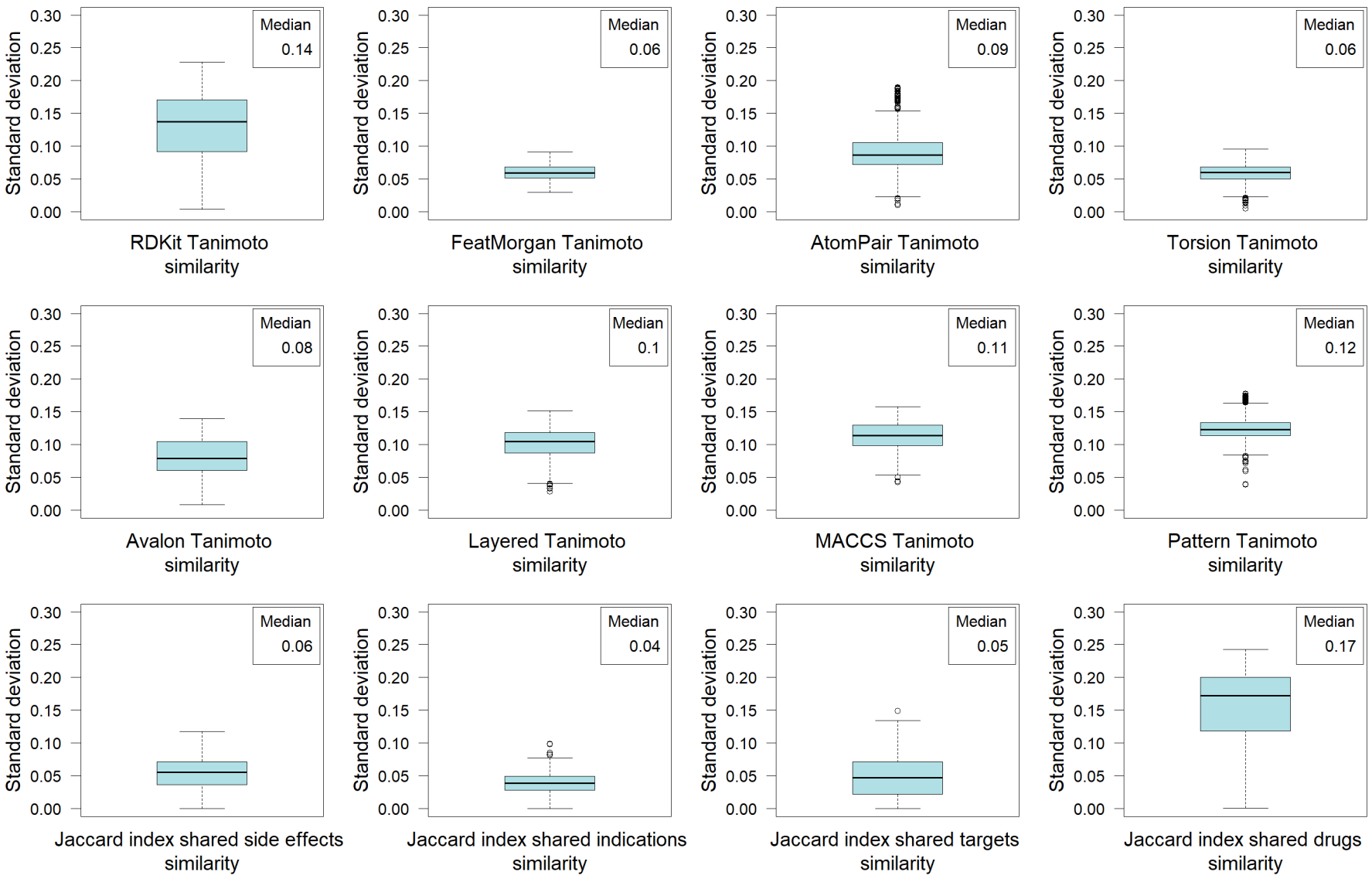


Fig. 4.27: The box-and-whisker plots showing the standard deviation distribution for each feature

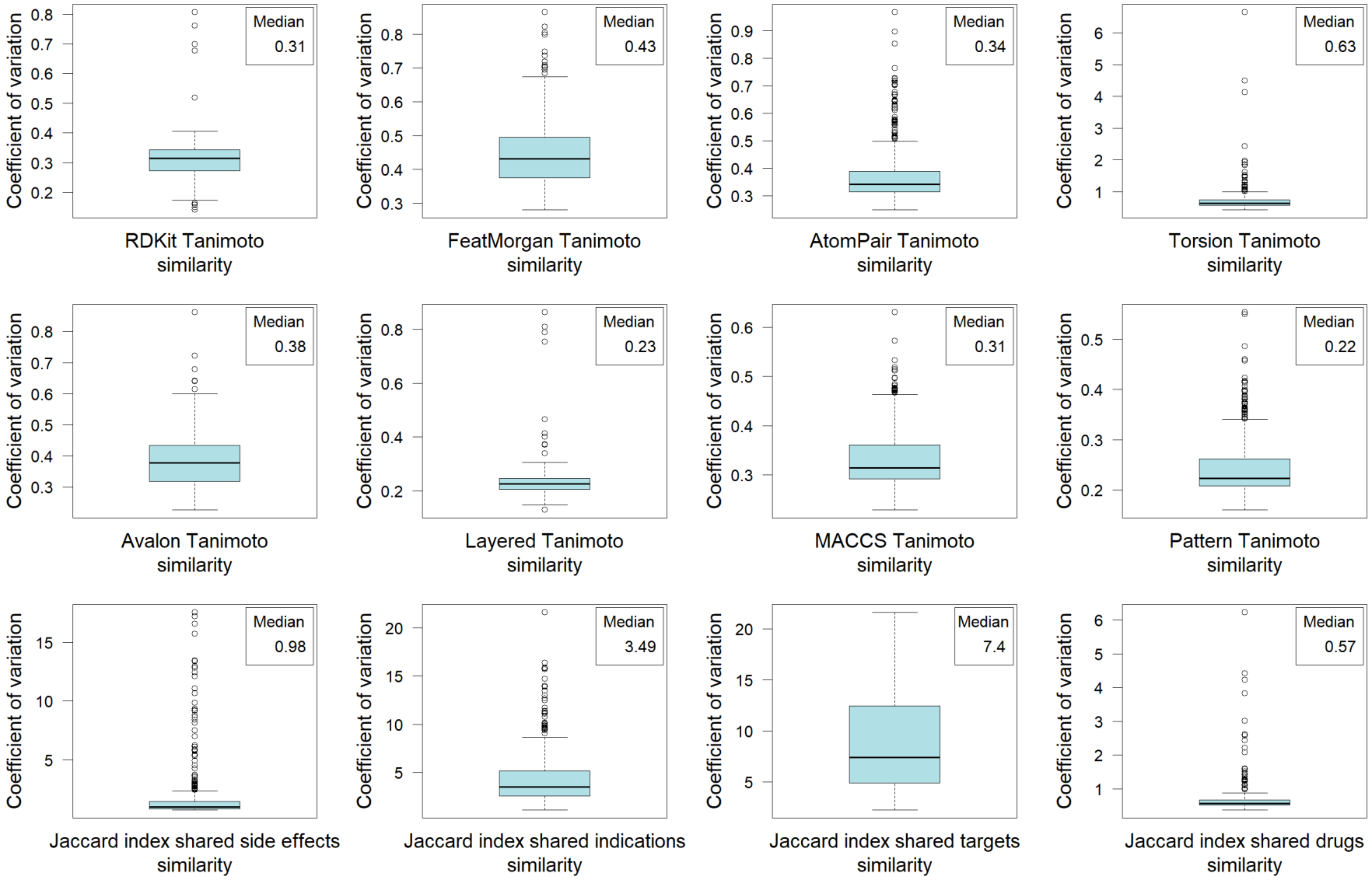


Fig. 4.28: The box-and-whisker plots showing the coefficient of the variation distribution for each feature. A coefficient of variation to be less than 1 corresponds to low-variance distributions.

4.4.2 Data distribution in groups

The data distribution in both groups (the group of examined positive and the group of negative association with a given side effect) can be plotted as box-and-whisker plots including strip charts for each feature via workflow below (Fig. **??**). The resulting box-and-whisker plots of the example dataset visualization suggest increased differences between both groups in selection datasets (Fig. 4.31).

Following on, a percentage difference was calculated for each feature in all datasets and the distribution of the results has been plotted as box-and-whisker plots, see the example below (Fig. 4.31). The box-and-whisker plots represent the features DBM/OVS ratios (%). However, as the percentage difference between both of the groups was not sufficient, additional filtering was applied to obtain features with a greater value difference (Fig. 4.29).

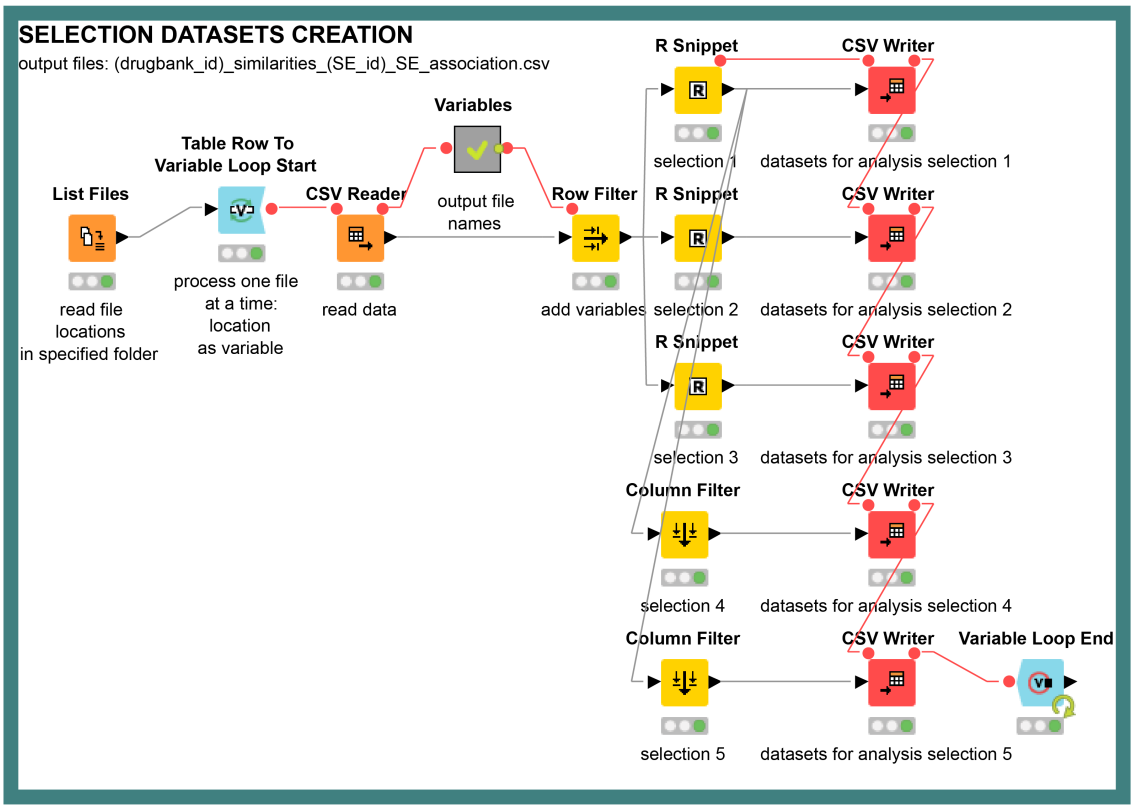


Fig. 4.29: The workflow for creating the selection datasets. Output data files are saved in specified locations set in 'Variables' metanode.

Although, the size of the datasets decreased the features are expected to have more predictive power. As the resulting box-and-whisker plots suggest (Fig. 4.32 and 4.33), the difference between both of the groups increased for each feature after this filtering step.

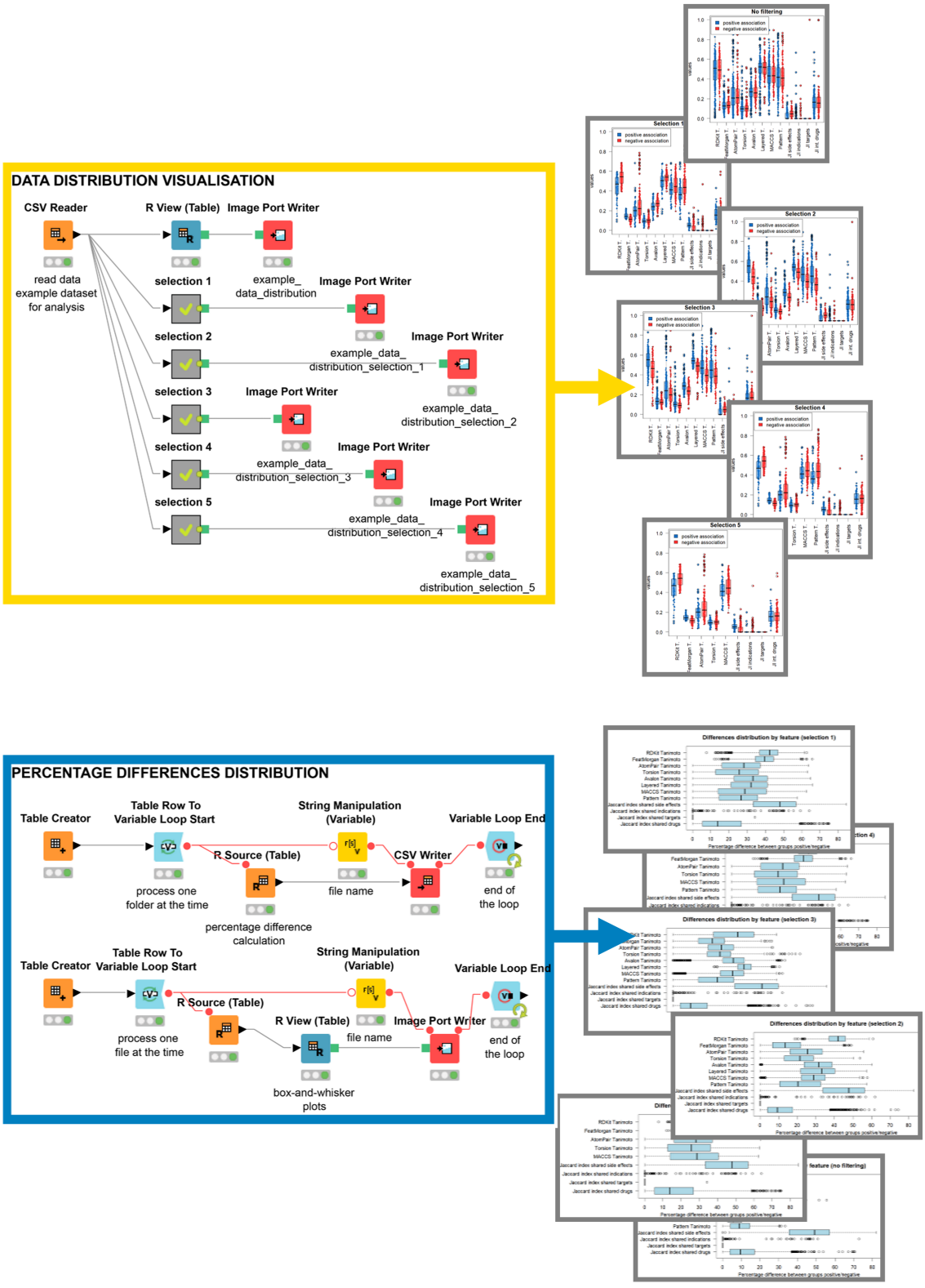


Fig. 4.30: The workflows for data distribution visualization per group and percentages differences distribution visualization. The results are visualized as box-and-whisker plots.

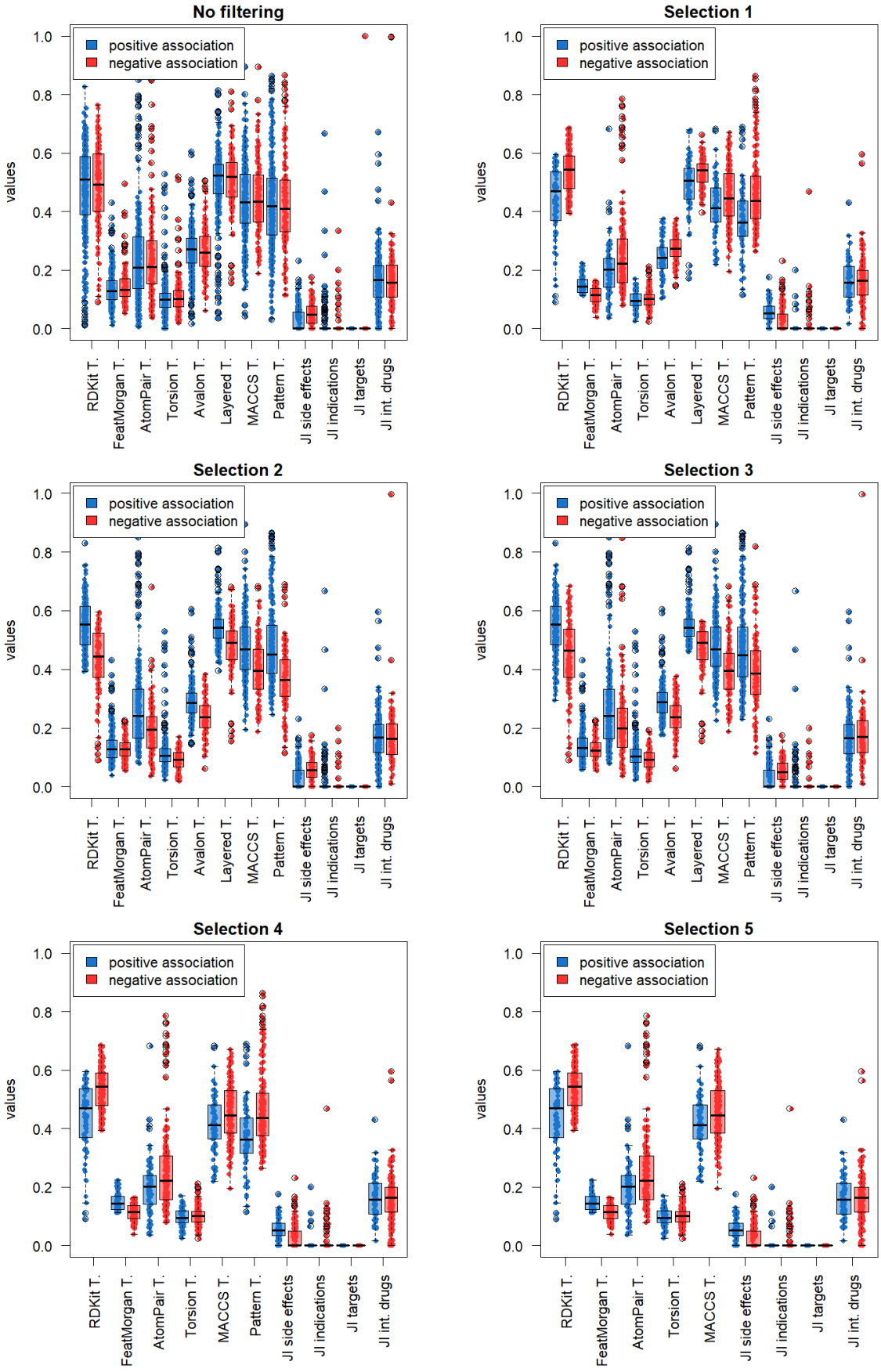


Fig. 4.31: The exemplary data distribution per feature by group (dataset for drug DB01195 and side effect C0013395; T = Tanimoto; JI = Jaccard index)

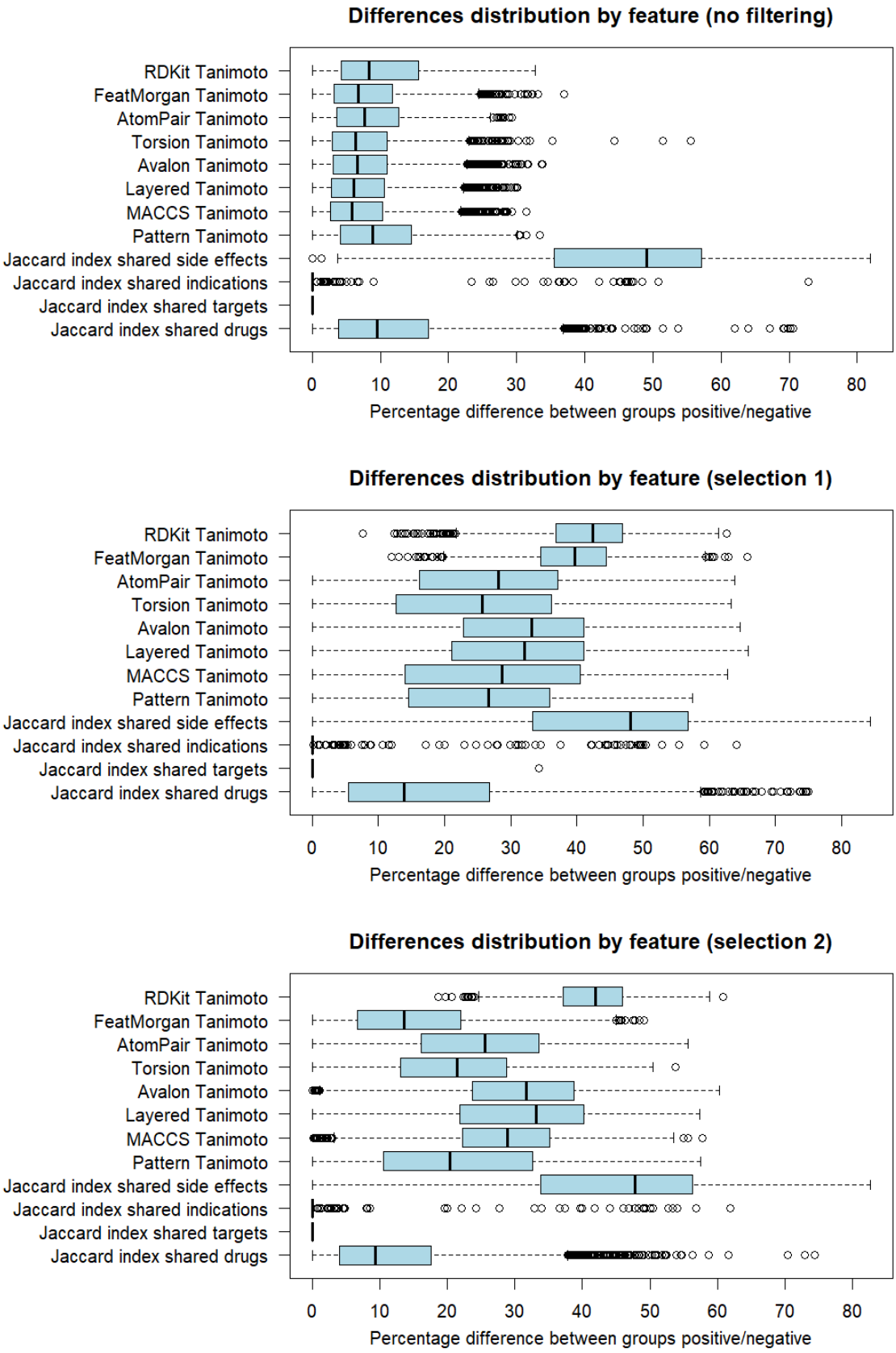


Fig. 4.32: The percentage differences distribution between the group of positives and the group of negatives for each feature I. Datasets: no filtering, selection 1, selection 2.

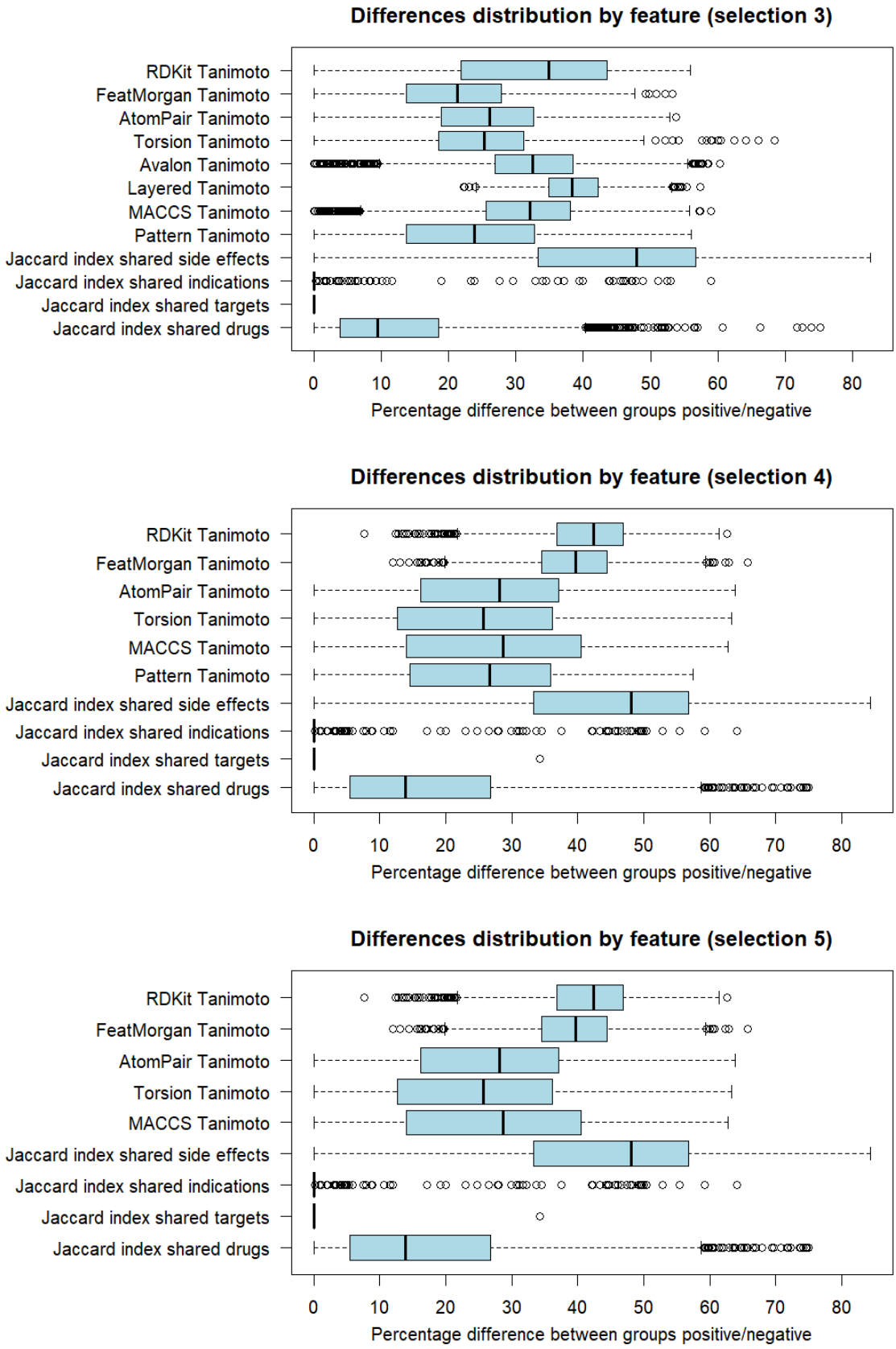


Fig. 4.33: The percentage differences distribution between the group of positives and the group of negatives for each feature I. Datasets: selection 3, selection 4, selection 5.

4.5 Model creation and evaluation

The third part of the set of the workflows deals with creating the models and evaluating their performance and predicting side effects and evaluating the predictions (Fig. 4.34).

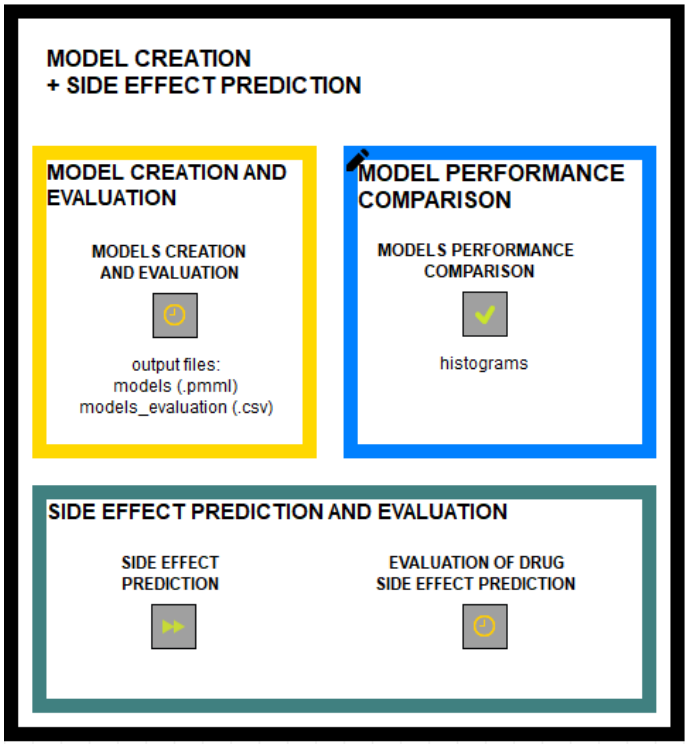


Fig. 4.34: The workflow overview - part III: a models creation, a models performance evaluation, a side effect prediction and a side effect predictions evaluation

The purpose of the following workflow is to train models in a loop to be later used for predicting drug side effects and save specific drug data table and model evaluation metrics (Fig. 4.35). The collected data are split into a training part and a 10-fold cross validation is applied using 'X-Partitioner' and 'X-Aggregator' KNIME nodes. After the data is partitioned, a decision tree model is trained using the training dataset in 'Decision Tree Learner' node and applied via 'Decision Tree Predictor' node on test dataset. In each iteration of the loop, only one drug and one side effect are analyzed at the same time.

Part of the workflow for creating the classification model (Fig. 4.35) contained in the 'Result table' metanode is intended for calculating model performance metrics. We evaluated each model performance using the 'Scorer' KNIME node which calculates several quality measures. It provides a confusion matrix, class prediction statistics and overall accuracy statistics. We studied the metrics of success via the accuracy statistics. In addition, the AUC-PR and the AUC-ROC values were

calculated via integrated R script in the 'R Snippet' KNIME node. All resulting values are appended in a final .csv file and visualized as described in the following subchapter.

To be able to transmit the model between the KNIME workflows or other data mining software, the model is stored via the 'PMML Writer' KNIME node in a standard XML-based portable PMML (Predictive Model Markup Language) format (.pmml) in each iteration. The PMML model can be then executed on new data. In the KNIME workflows the models can be executed via the 'PMML reader' and the 'PMML Compiling Predictor' KNIME nodes.

4.5.1 Model performance comparison

In total, 4,690 machine learning models were generated for 10 side effects and 469 examined drugs. Consequently, the model performance and side effect prediction were evaluated.

At first, box-and-whisker plots were constructed to display the distribution of the model performance metrics. The data for all the performance measures were processed and plotted in a loop via the integrated R scripts in the 'R Source (Table)' and the 'R View (Table)' KNIME nodes shown in the corresponding workflow figure (Fig. 4.36). The resulting plots of measures distributions were discussed further. We evaluated the performance of our predictors by their accuracy, AUC-ROC and AUC-PR values.

The accuracy statistics are shown in the figure below (Fig. 4.37). We can observe that the models performed higher overall accuracy when a filtering function was applied on the dataset. The best model accuracy was yielded by constructing the models with the 'Selection 1' dataset. The median value at which models reach a classification accuracy is around 84%. The model accuracy improved as it is shown in the figure.

However, as predicting the 'false' category was correct on more occasions, accuracy is a misleading metric. Therefore, true positives, false positives are regarded as key measures. As mentioned above, AUC is useful for comparing the performances of multiple learning methods. From the histograms of the AUC-PR (Fig. 4.37) and the AUC-ROC (Fig. 4.37) values, we can observe that the model performance improved for the models trained on datasets with a filtering function applied as higher AUC-PR values correspond to better model performance.

As our datasets are imbalanced and the number of negative outcomes is higher in all datasets, we selected AUC-PR as a more appropriate metric than AUC-ROC for the model comparison. AUC-PR is more sensitive to the improvements for the positive class than AUC-ROC is.

It is apparent that the range of all performance metrics of the models trained on datasets with no filtering is larger than of the others. The interquartile range of 'No filtering' models is larger in all performance metrics, meaning its performance metrics are less consistent around the median than those of the other models.

The data of the present study show that, combinations of different similarity coefficients improve classification results. All performance metrics of 'Selection 1' models have a higher maximum and median than others. The plots clearly show 'Selection 1' models classifiers outperform all other classifiers which is supported by all performance metrics showing a substantial increase. By observing the 'Selection 1' models results, we can conclude that additional features positively impact model performance.

4.5.2 Side effect prediction evaluation

The aim of the subsequent workflow (Fig. 4.38) is to predict side effects using the previously learned models. Each model is intended to test only a single given drug–side effect combination.

The workflow is designed to read multiple files in a given location to be processed in a loop. It loads a table of desired query drug–side effect combinations, reads the corresponding model in a specified folder and performs an analysis for each loop iteration. The 'PMML Reader' KNIME node is used to import the previously stored KNIME model. The workflow loops over, generates a prediction column for each query drug, and appends the results in a .csv file.

The final workflow represents evaluating the side effect predictions (Fig. 4.39). In the workflow, the predicted results are compared to real data obtained from SIDER. The number of false negatives, false positives, true negatives, true positives, as well as the total number of correctly and incorrectly predicted side effects are computed. The results are plotted in histograms.

We explored if there tends to be a difference between the prediction accuracy of different side effects. The results are plotted below in histograms displaying the distribution of side effect prediction accuracy. It is apparent from the plots that the best results were yielded by predicting side effects with ID C0917801 corresponding to insomnia. In total, 74.84% of predictions of this side effect were correct.

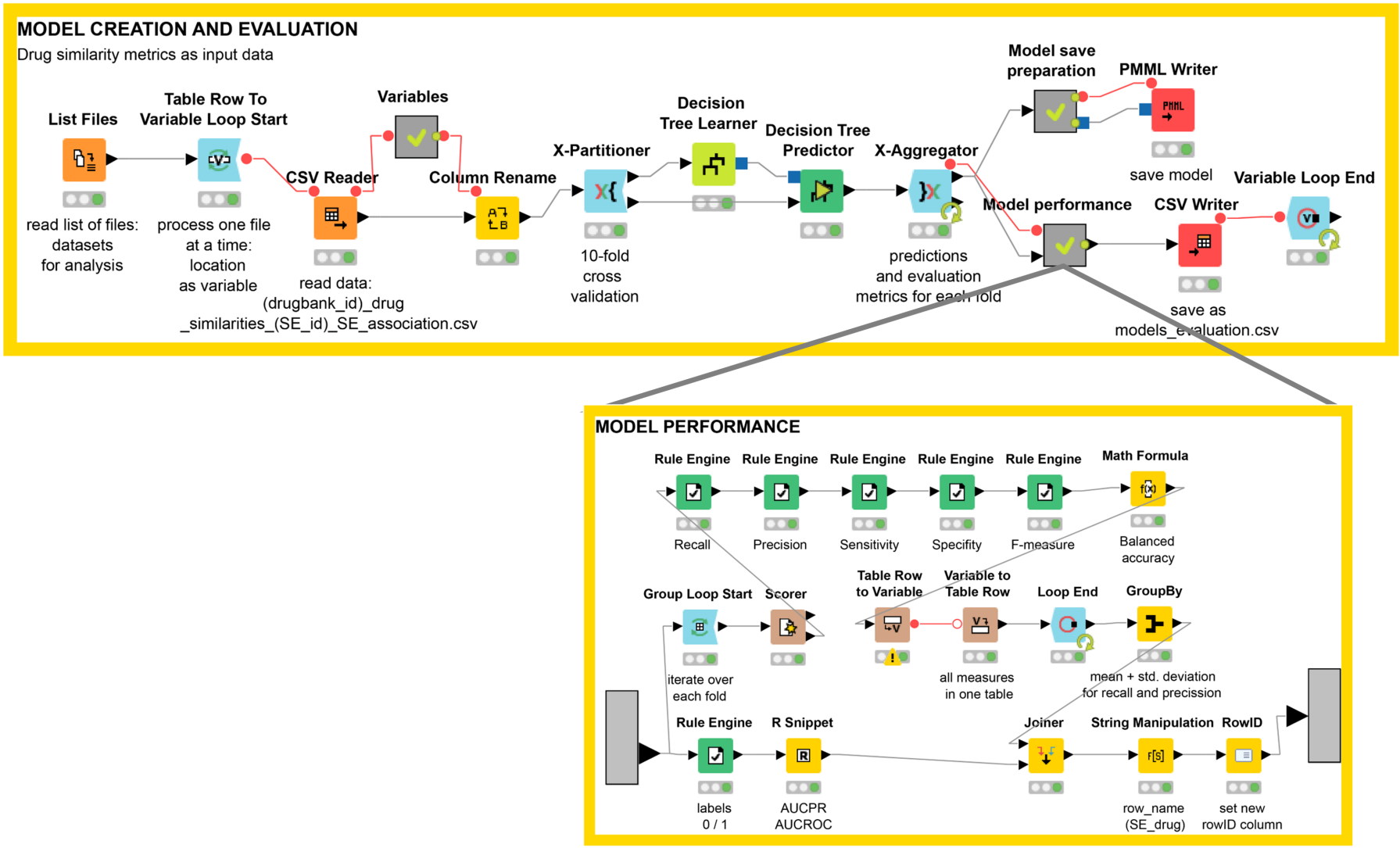


Fig. 4.35: The workflow for creating the models and evaluating their performance. The model performance metanode for calculating the model performance metrics is shown in detail at the bottom of the page.

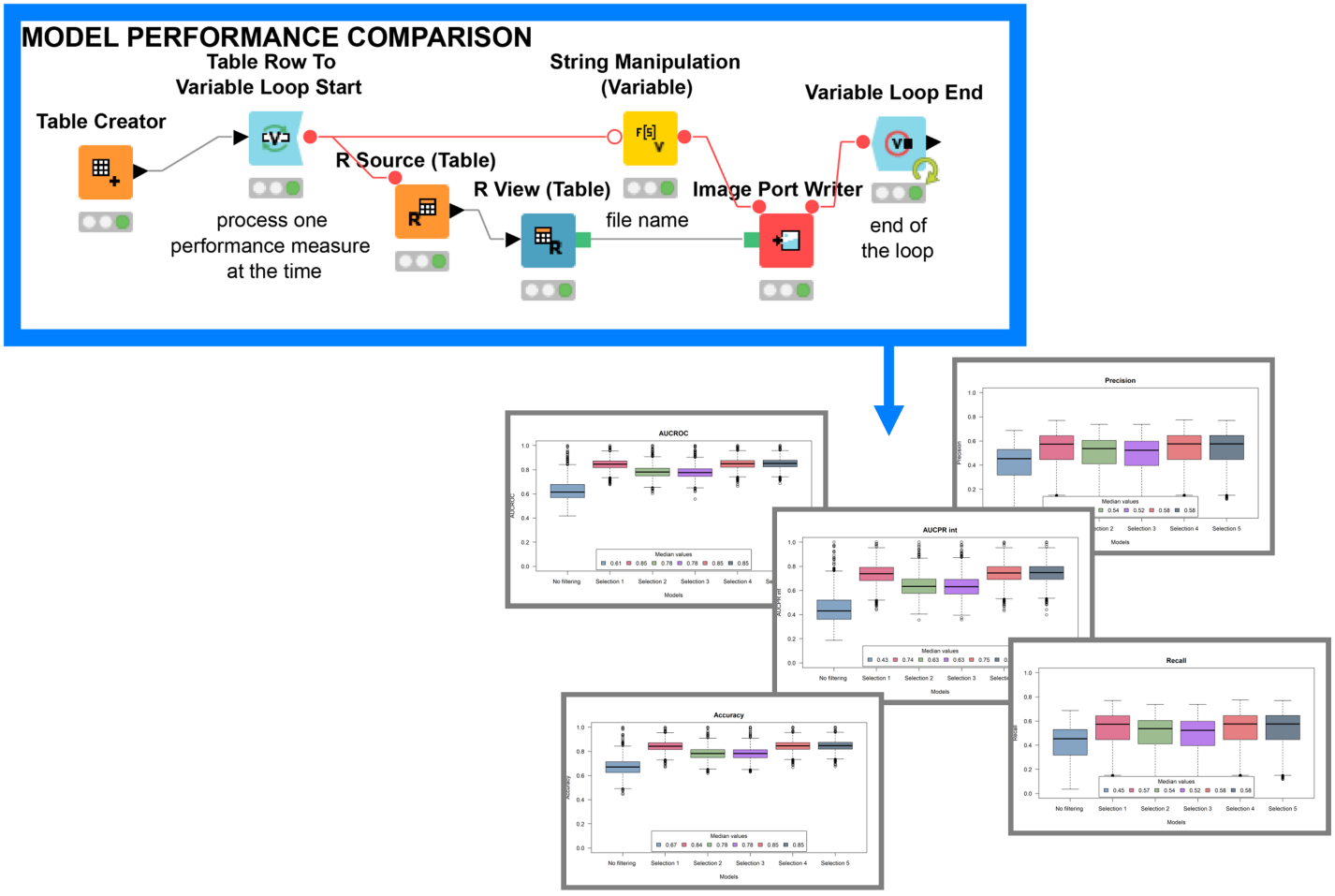


Fig. 4.36: The workflow for plotting metrics for the model performance comparison

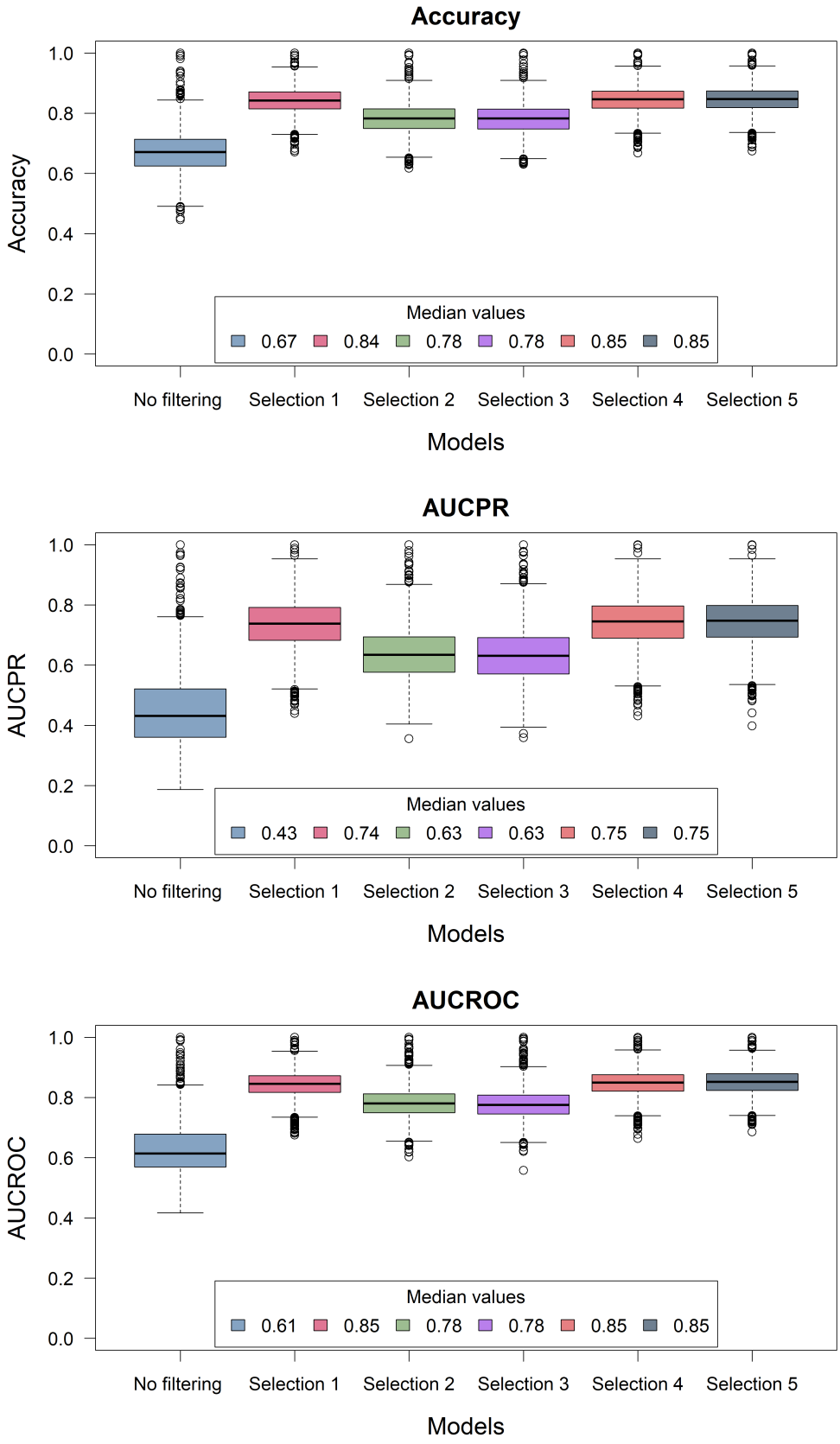


Fig. 4.37: The selected model performance measures. Models of Selection 1 datasets perform the best.

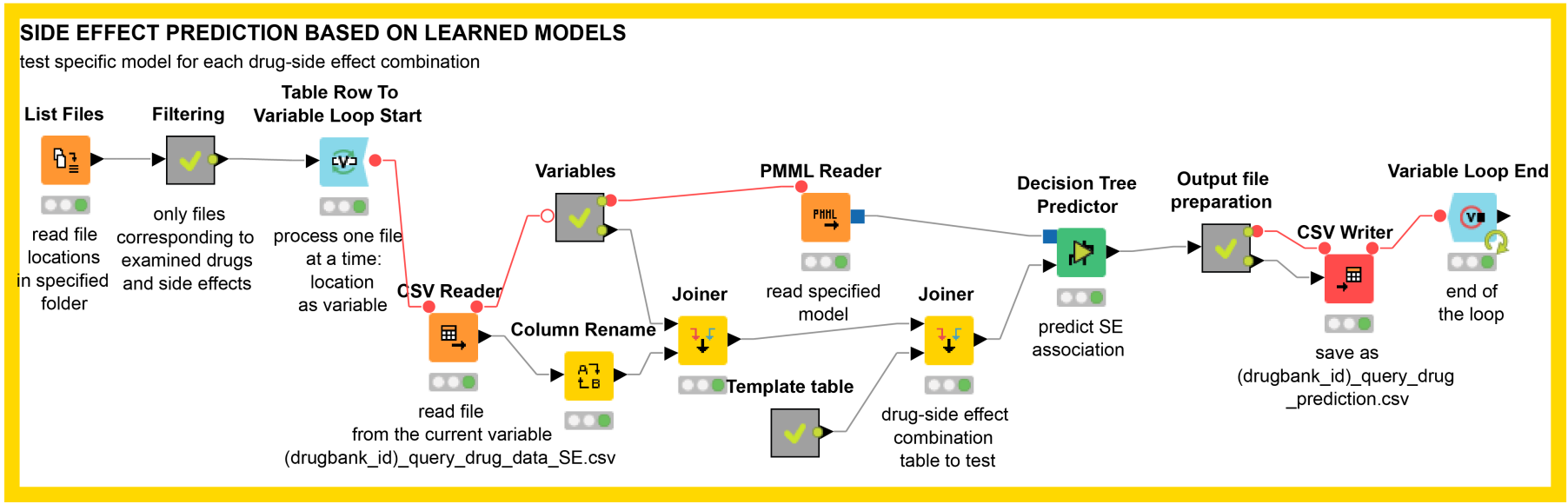


Fig. 4.38: The workflow for predicting side effects of a specific drug based on a specific learned model

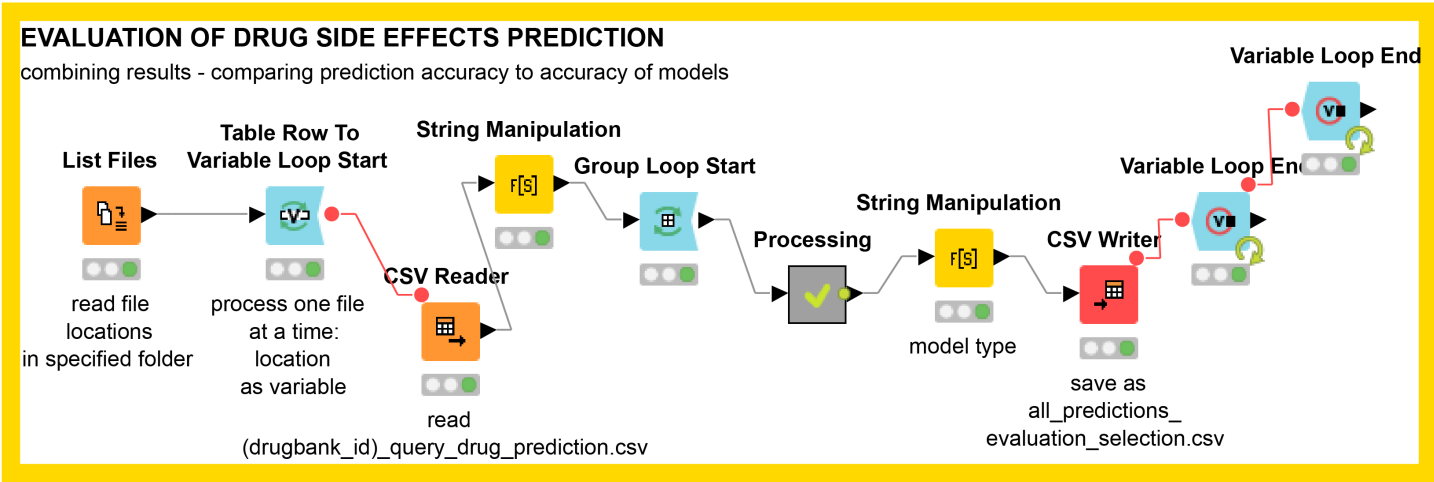


Fig. 4.39: The workflow for evaluating drug side effects predictions

5 Conclusion

Our work attempts to create a framework for predicting drug side effects. It was motivated by a lack of widely accepted standard protocols for drug–side effect data curation. Our workflow was supposed to incorporate data retrieval and exploration, machine-learning and an evaluation step for predicting side effects using a similarity-based learning method. The intention of our work to propose such a workflow has been completed. The hypothesis stating that drug molecules that share more similarities will demonstrate higher number of shared side effects compared to those drug molecules with fewer similarities has been proved to be correct.

The thesis provides *in silico* pipeline for predicting potential side effects. The proposed set of workflows implemented in open-source KNIME software is an easily accessible tool which should be relatively easy to understand by medicinal chemists and biomedical research scientists. The set of designed workflows integrate heterogeneous information into a single pipeline and allow for analyzing the integrated data in one place. Additionally, the workflows are able to provide an easy visual exploration which can be used in reports.

This work reached its proposed aim and achieved the main research objectives. The scientific background used in this work was described in theoretical and methods chapters. The proposed set of workflows was explained in detail in the Results and discussion chapter. We illustrated using the workflows with our filtered datasets of examined drugs and discussed the obtained results. Our calculations are based on FDA approved small molecule drug data – similarities between drug molecules in terms of structural features, shared side effects, shared targets, shared indications, and shared interacting drugs. The user can calculate other similarity features as needed in accordance with the proposed methodology.

That leads us to the major advantage of the designed set of workflows – its simple manipulation and re-usability. The tool enables customization to best suit the required specific needs of users without deeper programming abilities. Therefore, we hope to see applications of our set of workflows in future studies in which it can be combined with different approaches to build more complex methods. Such combinations of multiple data analysis could yield more relevant information and accelerate the drug development process and increase its efficiency.

Nevertheless, it is plausible that a number of limitations might have influenced the results obtained in our work. To begin with, the presented workflows can be extended by integrating other data sources. As we limit our analysis only to selected databases, further data collection from multiple other clinically relevant databases is required to obtain more reliable results.

Furthermore, our calculations are limited only to predicting a small number of

side effects as we studied only a set of the top ten most prevalent common and very common side effects. An extended analysis could focus on others. In our analysis, the best results have been achieved for predicting insomnia side effects. It has been demonstrated that prediction accuracy improved in models trained on a dataset in which a filtering function had been applied.

The limitations of this study also include the execution speed. A speed performance improvement can be reached by optimizing the configuration of the workflow nodes or using alternative nodes. Furthermore, the KNIME workflow may be improved by integrating Java scripts instead of R scripts, so that the data are processed in a more efficient way. Another way to speed up the execution of the workflows is running them on a big data cluster through specific KNIME extensions. Moreover, implementation of a proper error handling is needed.

The applied concept of similarity can be considered as another source of issue. Even though, similarity relationships can be problematic and the results of similarity based analyzes can lead to subjective interpretations, such methods are important complement to others drug development approaches.

Another direction to improve this study includes a better configuration of machine-learning models. E.g. the AUC-ROC curve was calculated by the trapezoidal rule, which simplifies calculating the area a lot. In order to get more accurate results, other calculations such as the Simpson or Romberg method could be integrated. In addition to the above, the accuracy of predicting the models could be boosted using ensemble learning methods which aggregate outputs from multiple models. This approach is used to reduce bias, decrease variance and achieve a better model performance. Also, comparisons to other machine learning algorithms would be beneficial. Advanced machine learning algorithms, including random forest, gradient boosting decision trees, deep neural networks could be used to enhance the performance of the prediction models. Also, more sophisticated statistical methods for datasets preparation and models performance analysis could be implemented in the workflow. More detailed analysis would be required to find out which of the applied similarity coefficients perform the best. It is needless to say that the biggest bottleneck of machine-learning-based approaches is that they cannot provide us with reliable results if the information about drugs, targets or interactions is missing.

In conclusion, despite all the limitations, we hope that our set of workflows will provide predictions allowing new research questions to be addressed and that the presented ideas will contribute to improving the efficiency of computational drug design. We believe that our developed tool provides a unique opportunity in predicting drug–side effect associations, and it is complementary to existing methods. Last but not least, it allows a broader portion of the scientific community to explore valuable data that is more and more available nowadays.

References

A.H.-L.; Y.G.S., 1992. Concepts and Applications of Molecular Similarity. *Journal of Molecular Structure*. Vol. 269, no. 3-4, pp. 376–377. ISSN 00222860. Available from DOI: 10.1016/0022-2860(92)85011-5.

AIN, Qurrat Ul; ALEKSANDROVA, Antoniya; ROESSLER, Florian D.; BALLESTER, Pedro J., 2015. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening: Machine-learning SFs to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. Vol. 5, no. 6, pp. 405–424. ISSN 17590876. Available from DOI: 10.1002/wcms.1225.

ALLEN, Benjamin C. P.; GRANT, Guy H.; RICHARDS, W. Graham, 2001. Similarity Calculations Using Two-Dimensional Molecular Representations. *Journal of Chemical Information and Computer Sciences*. Vol. 41, no. 2, pp. 330–337. ISSN 0095-2338. Available from DOI: 10.1021/ci0003956.

AMBURE, Pravin; BHAT, Jyotsna; PUZYN, Tomasz; ROY, Kunal, 2019. Identifying natural compounds as multi-target-directed ligands against Alzheimer's disease: an *in silico* approach. *Journal of Biomolecular Structure and Dynamics*. Vol. 37, no. 5, pp. 1282–1306. ISSN 0739-1102, ISSN 1538-0254. Available from DOI: 10.1080/07391102.2018.1456975.

ATIAS, Nir; SHARAN, Roded, 2011. An Algorithmic Framework for Predicting Side Effects of Drugs. *Journal of Computational Biology*. Vol. 18, no. 3, pp. 207–218. ISSN 1066-5277, ISSN 1557-8666. Available from DOI: 10.1089/cmb.2010.0255.

BAGHERIAN, Maryam; SABETI, Elyas; WANG, Kai; SARTOR, Maureen A; NIKOLOVSKA-COLESKA, Zaneta; NAJARIAN, Kayvan, 2020. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in Bioinformatics*. ISSN 1467-5463, ISSN 1477-4054. Available from DOI: 10.1093/bib/bbz157.

BAJORATH, Jürgen; OVERINGTON, John; JENKINS, Jeremy L; WALTERS, Pat, 2016. Drug discovery and development in the era of Big Data. *Future Medicinal Chemistry*. Vol. 8, no. 15, pp. 1807–1813. ISSN 1756-8919, ISSN 1756-8927. Available from DOI: 10.4155/fmc-2014-0081.

BAJUSZ, Dávid; RÁCZ, Anita; HÉBERGER, Károly, 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*. Vol. 7, no. 1. ISSN 1758-2946. Available from DOI: 10.1186/s13321-015-0069-3.

BEGAM, B.Firdaus; KUMAR, J. Satheesh, 2012. A Study on Cheminformatics and its Applications on Modern Drug Discovery. *Procedia Engineering*. Vol. 38, pp. 1264–1275. ISSN 18777058. Available from DOI: 10.1016/j.proeng.2012.06.156.

BEISKEN, Stephan; MEINL, Thorsten; WISWEDEL, Bernd; FIGUEIREDO, Luis F. de; BERTHOLD, Michael; STEINBECK, Christoph, 2013. KNIME-CDK: Workflow-driven cheminformatics. *BMC bioinformatics*. Vol. 14, no. 1, p. 257.

BENDER, Andreas; GLEN, Robert C., 2004. Molecular similarity: a key technique in molecular informatics. *Organic & Biomolecular Chemistry*. Vol. 2, no. 22, p. 3204. ISSN 1477-0520, ISSN 1477-0539. Available from DOI: 10.1039/b409813g.

BENTON, Wade W.; BROTHERS, Adam W.; JEFFERIS KIRK, Christa C.; LINGGI IRBY, Gretchen A.; RUBINO, Christopher M., 2011. Adverse Drug Reactions and Drug-drug Interactions. In: *Pediatric Critical Care*. Elsevier, pp. 1569–1589. ISBN 978-0-323-07307-3. Available from DOI: 10.1016/B978-0-323-07307-3.10118-1.

BERLIN, Jesse A.; GLASSER, Susan C.; ELLENBERG, Susan S., 2008. Adverse Event Detection in Drug Development: Recommendations and Obligations Beyond Phase 3. *American Journal of Public Health*. Vol. 98, no. 8, pp. 1366–1371. ISSN 0090-0036, ISSN 1541-0048. Available from DOI: 10.2105/AJPH.2007.124537.

BERNSTEIN, Frances C.; KOETZLE, Thomas F.; WILLIAMS, Graheme JB; MEYER, Edgar F.; BRICE, Michael D.; RODGERS, John R.; KENNARD, Olga; SHIMANOUCHI, Takehiko; TASUMI, Mitsuo, 1977. The protein data bank. A computer-based archival file for macromolecular structures. *European Journal of Biochemistry*. Vol. 80, no. 2, pp. 319–324. Available from DOI: 10.1111/j.1432-1033.1977.tb11885.x.

BERTHOLD, Michael R.; CEBRON, Nicolas; DILL, Fabian; GABRIEL, Thomas R.; KÖTTER, Tobias; MEINL, Thorsten; OHL, Peter; SIEB, Christoph; THIEL, Kilian; WISWEDEL, Bernd, 2007. KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer. ISBN 978-3-540-78239-1.

BIOVIA, Dassault Systèmes, [n.d.]. *Pipeline Pilot*. Available also from: http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/.

BLOWER, Paul; CROSS, Kevin, 2006. Decision Tree Methods in Pharmaceutical Research. *Current Topics in Medicinal Chemistry*. Vol. 6, no. 1, pp. 31–39. ISSN 15680266. Available from DOI: 10.2174/156802606775193301.

BRESSO, Emmanuel; GRISONI, Renaud; MARCHETTI, Gino; KARABOGA, Arnaud Sinan; SOUCHET, Michel; DEVIGNES, Marie-Dominique; SMAÏL-TABBONE, Malika, 2013. Integrative relational machine-learning for understanding drug side-effect profiles. *BMC Bioinformatics*, p. 12.

BROWN, Elliot G.; WOOD, Louise; WOOD, Sue, 1999. The Medical Dictionary for Regulatory Activities (MedDRA): *Drug Safety*. Vol. 20, no. 2, pp. 109–117. ISSN 0114-5916. Available from DOI: 10.2165/00002018-199920020-00002.

CARHART, Raymond E.; SMITH, Dennis H.; VENKATARAGHAVAN, R., 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*. Vol. 25, no. 2, pp. 64–73. ISSN 0095-2338. Available from DOI: 10.1021/ci00046a002.

CERETO-MASSAGUÉ, Adrià; OJEDA, María José; VALLS, Cristina; MULERO, Miquel; GARCIA-VALLVÉ, Santiago; PUJADAS, Gerard, 2015. Molecular fingerprint similarity search in virtual screening. *Methods*. Vol. 71, pp. 58–63. ISSN 10462023. Available from DOI: 10.1016/j.ymeth.2014.08.005.

CHAST, François, 2008. Chapter 1 - A History of Drug Discovery: From first steps of chemistry to achievements in molecular pharmacology. In: WERMUTH, Camille Georges (ed.). *The Practice of Medicinal Chemistry (Third Edition)*. Third Edition. New York: Academic Press, pp. 1–62. ISBN 978-0-12-374194-3. Available from DOI: 10.1016/B978-0-12-374194-3.00001-9.

ChemAxon, [n.d.]. Available also from: https://www.chemaxon.com.

ChemSpider, [n.d.]. Available also from: http://www.chemspider.com.

CHEN, Xi; LIU, Ming; GILSON, Michael K., 2001. BindingDB: a web-accessible molecular recognition database. *Combinatorial chemistry & high throughput screening*. Vol. 4, no. 8, pp. 719–725.

CHEN, Xing; REN, Biao; CHEN, Ming; WANG, Quanxin; ZHANG, Lixin; YAN, Guiying, 2016. NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. *PLOS Computational Biology*. Vol. 12, no. 7, e1004975. ISSN 1553-7358. Available from DOI: 10.1371/journal.pcbi.1004975.

CIOMS, 1995. *Guidelines for Preparing Core Clinical-safety Information on Drugs: Report of CIOMS Working Group III*. Geneva: Council for International Organizations of Medical Sciences. ISBN 978-92-9036-062-9.

COSTA, Pedro R; ACENCIO, Marcio L; LEMKE, Ney, 2010. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics*. Vol. 11, no. Suppl 5, S9. ISSN 1471-2164. Available from DOI: 10.1186/1471-2164-11-S5-S9.

DAVIS, Jesse; GOADRICH, Mark, 2006. The relationship between Precision-Recall and ROC curves. In: Pittsburgh, Pennsylvania: ACM Press, pp. 233–240. ISBN 978-1-59593-383-6. Available from DOI: 10.1145/1143844.1143874.

DEMŠAR, Janez; CURK, Tomaž; ERJAVEC, Aleš; GORUP, Črt; HOČEVAR, Tomaž; MILUTINOVIČ, Mitar; MOŽINA, Martin; POLAJNAR, Matija; TOPLAK, Marko; STARIČ, Anže, 2013. Orange: data mining toolbox in Python. *The Journal of Machine Learning Research*. Vol. 14, no. 1, pp. 2349–2353.

DIMITRI, Giovanna Maria; LIÓ, Pietro, 2017. DrugClust: A machine learning approach for drugs side effects prediction. *Computational Biology and Chemistry*. Vol. 68, pp. 204–210. ISSN 14769271. Available from DOI: 10.1016/j.compbiolchem.2017.03.008.

DING, Yijie; TANG, Jijun; GUO, Fei, 2019. Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing*. Vol. 325, pp. 211–224. ISSN 09252312. Available from DOI: 10.1016/j.neucom.2018.10.028.

DITOMMASO, Jack Antonio, 2017. Implementation of virtual workflows in KNIME for medicinal chemistry. *Journal of Student Science and Technology*. Vol. 10, no. 1. ISSN 1913-1925, ISSN 1913-1925. Available from DOI: 10.13034/jsst.v10i1.123.

DOWDEN, Helen; MUNRO, Jamie, 2019. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery*. Vol. 18, no. 7, pp. 495–496. ISSN 1474-1776, ISSN 1474-1784. Available from DOI: 10.1038/d41573-019-00074-z.

DREWS, Jürgen, 2000. Drug discovery: a historical perspective. *Science*. Vol. 287, no. 5460, pp. 1960–1964.

DURANT, Joseph L.; LELAND, Burton A.; HENRY, Douglas R.; NOURSE, James G., 2002. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences*. Vol. 42, no. 6, pp. 1273–1280. ISSN 0095-2338. Available from DOI: 10.1021/ci010132r.

FALCÓN-CANO, Gabriela; MOLINA, Christophe; CABRERA-PÉREZ, Miguel Ángel, 2020. ADME Prediction with KNIME: Development and Validation of a Publicly Available Workflow for the Prediction of Human Oral Bioavailability. *Journal of Chemical Information and Modeling*. Vol. 60, no. 6, pp. 2660–2667. ISSN 1549-9596, ISSN 1549-960X. Available from DOI: 10.1021/acs.jcim.0c00019.

FILLBRUNN, Alexander; DIETZ, Christian; PFEUFFER, Julianus; RAHN, René; LANDRUM, Gregory A.; BERTHOLD, Michael R., 2017. KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology*. Vol. 261, pp. 149–156. ISSN 01681656. Available from DOI: 10.1016/j.jbiotec.2017.07.028.

GALLY, José-Manuel; BOURG, Stéphane; DO, Quoc-Tuan; ACI-SÈCHE, Samia; BONNET, Pascal, 2017. VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. *Molecular Informatics*. Vol. 36, no. 10, p. 1700023. ISSN 18681743. Available from DOI: 10.1002/minf.201700023.

GAO, Qingzhi; YANG, Lulu; ZHU, Yongqiang, 2010. Pharmacophore based drug design approach as a practical process in drug discovery. *Current computer-aided drug design*. Vol. 6, no. 1, pp. 37–49.

GAULTON, A.; BELLIS, L. J.; BENTO, A. P.; CHAMBERS, J.; DAVIES, M.; HERSEY, A.; LIGHT, Y.; MCGLINCHEY, S.; MICHALOVICH, D.; AL-LAZIKANI, B.; OVERINGTON, J. P., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*. Vol. 40, no. D1, pp. D1100–D1107. ISSN 0305-1048, ISSN 1362-4962. Available from DOI: 10.1093/nar/gkr777.

GEORGE, Rija; THOMAS, Sneha; JACOB, Sarah; GEORRGE, John J, 2017. Approaches for Novel Drug Target Identification, p. 24.

GERTRUDES, J. C.; MALTAROLLO, V. G.; SILVA, R. A.; OLIVEIRA, P. R.; HONORIO, K. M.; DA SILVA, A. B. F., 2012. Machine learning techniques and drug design. *Current medicinal chemistry*. Vol. 19, no. 25, pp. 4289–4297.

GOOD, Andrew C.; RICHARDS, W.Graham, 2002. Explicit Calculation of 3D Molecular Similarity. In: KUBINYI, Hugo; FOLKERS, Gerd; MARTIN, Yvonne C. (eds.). *3D QSAR in Drug Design*. Dordrecht: Springer Netherlands, pp. 321–338. ISBN 978-0-7923-4790-3. Available from DOI: 10.1007/0-306-46857-3_17.

GOODNOW, Robert A., 2006. Hit and lead identification: Integrated technology-based approaches. *Drug Discovery Today: Technologies*. Vol. 3, no. 4, pp. 367–375. ISSN 17406749. Available from DOI: 10.1016/j.ddtec.2006.12.009.

GRAU, Jan; GROSSE, Ivo; KEILWAGEN, Jens, 2015. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*. Vol. 31, no. 15, pp. 2595–2597.

GUNTHER, S.; KUHN, M.; DUNKEL, M.; CAMPILLOS, M.; SENGER, C.; PETSALAKI, E.; AHMED, J.; URDIALES, E. G.; GEWIESS, A.; JENSEN, L. J.; SCHNEIDER, R.; SKOBLO, R.; RUSSELL, R. B.; BOURNE, P. E.; BORK, P.; PREISSNER, R., 2007. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Research*. Vol. 36, no. Database, pp. D919–D922. ISSN 0305-1048, ISSN 1362-4962. Available from DOI: 10.1093/nar/gkm862.

GÜTLEIN, Martin; KARWATH, Andreas; KRAMER, Stefan, 2014. CheS-Mapper 2.0 for visual validation of (Q) SAR models. *Journal of cheminformatics*. Vol. 6, no. 1, p. 41.

HAUPT, V. Joachim; DAMINELLI, Simone; SCHROEDER, Michael, 2013. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE*. Vol. 8, no. 6, e65894. ISSN 1932-6203. Available from DOI: 10.1371/journal.pone.0065894.

HUGHES, Jp; REES, S; KALINDJIAN, Sb; PHILPOTT, Kl, 2011. Principles of early drug discovery: Principles of early drug discovery. *British Journal of Pharmacology*. Vol. 162, no. 6, pp. 1239–1249. ISSN 00071188. Available from DOI: 10.1111/j.1476-5381.2010.01127.x.

IRWIN, John J.; SHOICHET, Brian K., 2005. ZINC-a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*. Vol. 45, no. 1, pp. 177–182.

KATSILA, Theodora; SPYROULIAS, Georgios A.; PATRINOS, George P.; MATSOUKAS, Minos-Timotheos, 2016. Computational approaches in target identification and drug discovery. *Computational and Structural Biotechnology Journal*. Vol. 14, pp. 177–184. ISSN 20010370. Available from DOI: 10.1016/j.csbj.2016.04.004.

KESERŰ, György M.; MAKARA, Gergely M., 2006. Hit discovery and hit-to-lead approaches. *Drug Discovery Today*. Vol. 11, no. 15-16, pp. 741–748. ISSN 13596446. Available from DOI: 10.1016/j.drudis.2006.06.016.

KHAMIS, Mohamed A.; GOMAA, Walid; AHMED, Walaa F., 2015. Machine learning in computational docking. *Artificial Intelligence in Medicine*. Vol. 63, no. 3, pp. 135–152. ISSN 09333657. Available from DOI: 10.1016/j.artmed.2015.02.002.

KHAN, Muhammad Irfan, 2017. *Drug side-effect prediction using machine learning methods*. Computer Science. Available also from: https://aaltodoc.aalto.fi/handle/123456789/29305.

KNOWLES, Jonathan; GROMO, Gianni, 2003. A Guide to Drug Discovery: Target selection in drug discovery. *Nature Reviews Drug Discovery*. Vol. 2, no. 1, pp. 63–69. ISSN 14741776, ISSN 14741784. Available from DOI: 10.1038/nrd986.

KOŠČOVÁ, P.; PROVAZNÍK, I., 2016. Pharmacophore modelling used in rational drug design. *Chemické Listy*. Vol. 110, no. 8, pp. 575–580. ISSN 0009-2770.

KUHN, Michael; LETUNIC, Ivica; JENSEN, Lars Juhl; BORK, Peer, 2016. The SIDER database of drugs and side effects. *Nucleic Acids Research*. Vol. 44, no. D1, pp. D1075–D1079. ISSN 0305-1048, ISSN 1362-4962. Available from DOI: 10.1093/nar/gkv1075.

LANDRUM, Greg, [n.d.]. RDKit Documentation, p. 159.

LAVECCHIA, Antonio, 2015. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*. Vol. 20, no. 3, pp. 318–331. ISSN 13596446. Available from DOI: 10.1016/j.drudis.2014.10.012.

LEE, Man-Ling; ALIAGAS, Ignacio; FENG, Jianwen A.; GABRIEL, Thomas; O'DONNELL, T. J.; SELLERS, Benjamin D.; WISWEDEL, Bernd; GOBBI, Alberto, 2017. chemalot and chemalot_knime: Command line programs as workflow tools for drug discovery. *Journal of Cheminformatics*. Vol. 9, no. 1. ISSN 1758-2946. Available from DOI: 10.1186/s13321-017-0228-9.

LI, Ying Hong; YU, Chun Yan; LI, Xiao Xu; ZHANG, Peng; TANG, Jing; YANG, Qingxia; FU, Tingting; ZHANG, Xiaoyu; CUI, Xuejiao; TU, Gao; ZHANG, Yang; LI, Shuang; YANG, Fengyuan; SUN, Qiu; QIN, Chu; ZENG, Xian; CHEN, Zhe; CHEN, Yu Zong; ZHU, Feng, 2018. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Research*. Vol. 46, no. D1, pp. D1121–D1127. ISSN 1362-4962. Available from DOI: 10.1093/nar/gkx1076.

LI, Zhan-Chao; HUANG, Meng-Hua; ZHONG, Wen-Qian; LIU, Zhi-Qing; XIE, Yun; DAI, Zong; ZOU, Xiao-Yong, 2016. Identification of drug–target interaction from interactome network with 'guilt-by-association' principle and topology features. *Bioinformatics*. Vol. 32, no. 7, pp. 1057–1064. ISSN 1367-4803, ISSN 1460-2059. Available from DOI: 10.1093/bioinformatics/btv695.

LIANG, Haiyan; CHEN, Lei; ZHAO, Xian; ZHANG, Xiaolin, 2020. Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy. *Computational and Mathematical Methods in Medicine*. Vol. 2020, pp. 1–16. ISSN 1748-670X, ISSN 1748-6718. Available from DOI: 10.1155/2020/1573543.

LIU, T.; LIN, Y.; WEN, X.; JORISSEN, R. N.; GILSON, M. K., 2007. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*. Vol. 35, no. Database, pp. D198–D201. ISSN 0305-1048, ISSN 1362-4962. Available from DOI: 10.1093/nar/gkl999.

M., Ali; A., Ezzat, 2020. *DrugBank Database XML Parser. Dainanahan. R package version 1.2.0, https://CRAN.R-project.org/package=dbparser*.

MAGGIORA, Gerald; VOGT, Martin; STUMPFE, Dagmar; BAJORATH, Jürgen, 2014. Molecular Similarity in Medicinal Chemistry: Miniperspective. *Journal of Medicinal Chemistry*. Vol. 57, no. 8, pp. 3186–3204. ISSN 0022-2623, ISSN 1520-4804. Available from DOI: 10.1021/jm401411z.

MARTIN, Yvonne C.; KOFRON, James L.; TRAPHAGEN, Linda M., 2002. Do Structurally Similar Molecules Have Similar Biological Activity? *Journal of Medicinal Chemistry*. Vol. 45, no. 19, pp. 4350–4358. ISSN 0022-2623, ISSN 1520-4804. Available from DOI: 10.1021/jm020155c.

MATTHEWS, Holly; HANISON, James; NIRMALAN, Niroshini, 2016. "Omics"-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. *Proteomes*. Vol. 4, no. 3, p. 28. ISSN 2227-7382. Available from DOI: 10.3390/proteomes4030028.

MCGUIRE, Ross; VERHOEVEN, Stefan; VASS, Márton; VRIEND, Gerrit; ESCH, Iwan J. P. de; LUSHER, Scott J.; LEURS, Rob; RIDDER, Lars; KOOISTRA, Albert J.; RITSCHEL, Tina; GRAAF, Chris de, 2017. 3D-e-Chem-VM: Structural Cheminformatics Research Infrastructure in a Freely Available Virtual Machine. *Journal of Chemical Information and Modeling*. Vol. 57, no. 2, pp. 115–121. ISSN 1549-9596, ISSN 1549-960X. Available from DOI: 10.1021/acs.jcim.6b00686.

MITCHELL, Tom M., 1997. *Machine Learning*. New York: McGraw-Hill. McGraw-Hill series in computer science. ISBN 978-0-07-042807-2.

MOHD ALI, Yousoff Effendy; KWA, Kiam Heong; RATNAVELU, Kurunathan, 2017. Predicting new drug indications from network analysis. *International Journal of Modern Physics C*. Vol. 28, no. 09, p. 1750118. ISSN 0129-1831, ISSN 1793-6586. Available from DOI: 10.1142/S0129183117501182.

Molecular Operating Environment (MOE), 2013.08, 2017. Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7.

MYERS, Richard L., 2007. *The 100 most important chemical compounds: a reference guide*. Westport, Conn: Greenwood Press. ISBN 978-0-313-33758-1.

NICOLA, George; BERTHOLD, Michael R.; HEDRICK, Michael P.; GILSON, Michael K., 2015. Connecting proteins with drug-like compounds: Open source drug discovery workflows with BindingDB and KNIME. *Database*. Vol. 2015, bav087. ISSN 1758-0463. Available from DOI: 10.1093/database/bav087.

NILAKANTAN, Ramaswamy; BAUMAN, Norman; DIXON, J. Scott; VENKATARAGHAVAN, R., 1987. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*. Vol. 27, no. 2, pp. 82–85. ISSN 0095-2338. Available from DOI: 10.1021/ci00054a008.

NIU, Yanqing; ZHANG, Wen, 2017. Quantitative prediction of drug side effects based on drug-related features. *Interdisciplinary Sciences: Computational Life Sciences*. Vol. 9, no. 3, pp. 434–444. ISSN 1913-2751, ISSN 1867-1462. Available from DOI: 10.1007/s12539-017-0236-5.

O'BOYLE, Noel M.; SAYLE, Roger A., 2016. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics*. Vol. 8, no. 1, p. 36. ISSN 1758-2946. Available from DOI: 10.1186/s13321-016-0148-0.

PAPADATOS, George; DAVIES, Mark; DEDMAN, Nathan; CHAMBERS, Jon; GAULTON, Anna; SIDDLE, James; KOKS, Richard; IRVINE, Sean A.; PETTERSSON, Joe; GONCHAROFF, Nicko; HERSEY, Anne; OVERINGTON, John P., 2016. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research*. Vol. 44, pp. 1220–1228. Available from DOI: 10.1093/nar/gkv1253.

PAUWELS, Edouard; STOVEN, Véronique; YAMANISHI, Yoshihiro, 2011. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*. Vol. 12, no. 1, p. 169. ISSN 1471-2105. Available from DOI: 10.1186/1471-2105-12-169.

PLENGE, Robert M., 2016. Disciplined approach to drug discovery and early development. *Science Translational Medicine*. Vol. 8, no. 349, 349ps15–349ps15. ISSN 1946-6234, ISSN 1946-6242. Available from DOI: 10.1126/scitranslmed.aaf2608.

POLISHCHUK, P. G.; MADZHIDOV, T. I.; VARNEK, A., 2013. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design*. Vol. 27, no. 8, pp. 675–679. ISSN 0920-654X, ISSN 1573-4951. Available from DOI: 10.1007/s10822-013-9672-4.

RDKit: Open-source cheminformatics, [n.d.]. Available also from: http://www.rdkit.org.

ROUGHLEY, Stephen D., 2020. Five Years of the KNIME Vernalis Cheminformatics Community Contribution. *Current Medicinal Chemistry*. Vol. 27, no. 38, pp. 6495–6522. ISSN 09298673. Available from DOI: 10.2174/0929867325666180904113616.

SACHDEV, Kanica; GUPTA, Manoj K., 2020. A comprehensive review of computational techniques for the prediction of drug side effects. *Drug Development Research*. Vol. 81, no. 6, pp. 650–670. ISSN 0272-4391, ISSN 1098-2299. Available from DOI: 10.1002/ddr.21669.

SAITO, Takaya; REHMSMEIER, Marc, 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. Vol. 10, no. 3, e0118432. ISSN 1932-6203. Available from DOI: 10.1371/journal.pone.0118432.

SAKKIAH, Sugunadevi; THANGAPANDIAN, Sundarapandian; LEE, Keun Woo, 2012. Pharmacophore modeling, molecular docking, and molecular dynamics simulation approaches for identifying new lead compounds for inhibiting aldose reductase 2. *Journal of Molecular Modeling*. Vol. 18, no. 7, pp. 3267–3282. ISSN 1610-2940, ISSN 0948-5023. Available from DOI: 10.1007/s00894-011-1247-5.

SAUBERN, Simon; GUHA, Rajarshi; BAELL, Jonathan B., 2011. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Molecular Informatics*. Vol. 30, no. 10, pp. 847–850. ISSN 18681743. Available from DOI: 10.1002/minf.201100076.

SCHÄFER, Till; KRIEGE, Nils; HUMBECK, Lina; KLEIN, Karsten; KOCH, Oliver; MUTZEL, Petra, 2017. Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. *Journal of Cheminformatics*. Vol. 9, no. 1, p. 28. ISSN 1758-2946. Available from DOI: 10.1186/s13321-017-0213-3.

SCHEIBER, Josef; JENKINS, Jeremy L.; SUKURU, Sai Chetan K.; BENDER, Andreas; MIKHAILOV, Dmitri; MILIK, Mariusz; AZZAOUI, Kamal; WHITEBREAD, Steven; HAMON, Jacques; URBAN, Laszlo; GLICK, Meir; DAVIES, John W., 2009. Mapping Adverse Drug Reactions in Chemical Space. *Journal of Medicinal Chemistry*. Vol. 52, no. 9, pp. 3103–3107. ISSN 0022-2623, ISSN 1520-4804. Available from DOI: 10.1021/jm801546k.

SCHENONE, Monica; DANČÍK, Vlado; WAGNER, Bridget K; CLEMONS, Paul A, 2013. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology*. Vol. 9, no. 4, pp. 232–240. ISSN 1552-4450, ISSN 1552-4469. Available from DOI: 10.1038/nchembio.1199.

Schrödinger KNIME Extensions, 2021. Schrödinger, LLC, New York, NY.

SEO, Sukyung; LEE, Taekeon; KIM, Mi-hyun; YOON, Youngmi, 2020. Prediction of Side Effects Using Comprehensive Similarity Measures. *BioMed Research International*. Vol. 2020, pp. 1–10. ISSN 2314-6133, ISSN 2314-6141. Available from DOI: 10.1155/2020/1357630.

SHAFER, John; AGRAWAL, Rakeeh; MEHTA, Manish, 1996. SPRINT: A Scalable Parallel Classifier for Data Mining. *Proceedings of the 22nd VLDB Conference Mumbai(Bombay), India*, p. 12.

STEINBECK, Christoph; HOPPE, Christian; KUHN, Stefan; FLORIS, Matteo; GUHA, Rajarshi; WILLIGHAGEN, Egon, 2006. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Current Pharmaceutical Design*. Vol. 12, no. 17, pp. 2111–2120. ISSN 13816128. Available from DOI: 10.2174/138161206777585274.

STEVENS, Erland, 2016. Medicinal Chemistry: The Molecular Basis of Drug Discovery [MOOC]. *DavidsonX*. Available also from: https://www.edx.org/course/medicinal-chemistry-the-molecular-basis-of-drug-di.

STROBELT, Hendrik; BERTINI, Enrico; BRAUN, Joachim; DEUSSEN, Oliver; GROTH, Ulrich; MAYER, Thomas U.; MERHOF, Dorit, 2012. HiTSEE KNIME: a visualization tool for hit selection and analysis in high-throughput screening experiments for the KNIME platform. *BMC bioinformatics*. Vol. 13, no. 8, S4.

TEAM, R Development Core, 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available also from: http://www.R-project.org.

TIWARI, Abhishek; SEKHAR, Arvind K.T., 2007. Workflow based framework for life science informatics. *Computational Biology and Chemistry*. Vol. 31, no. 5-6, pp. 305–319. ISSN 14769271. Available from DOI: 10.1016/j.compbiolchem.2007.08.009.

TUBACH, Florence; LAMARQUE-GARNIER, Véronique; CASTOT, Anne; AUCLERT, Laurent; BONNIN, Marthe; DAUDIN, Magda; DUBOIS, Catherine; FRANCILLON, Alain; FRAUGER, Elisabeth; GIRAULT, Danièle; GOURLAY, Marie-Laurence; JOLLIET, Pascale; JAÏS, Carmen Kreft; GRIMALDI, Lamiae; LIÈVRE, Michel; MAILLÈRE, Patricia; MAUGENDRE, Philippe; MICALLEF, Joelle; MIRANDA, Sara; PARIENTE, Antoine; BIGOT, Sylvie Paulmier; PENTEL, Jonathan; PRESTAT, Laure; PRUVOT, Fanny; FERRER, Valérie Querol; ROCHER, Fanny; SAUSSIER, Christel; ZANETTI, Laura, 2011. Role of the Post-Marketing Authorisation Studies in Drug Risk Surveillance: Specifications and Methodologies. *Therapies*. Vol. 66, no. 4, pp. 355–362. ISSN 00405957. Available from DOI: 10.2515/therapie/2011048.

TUERKOVA, Alzbeta; ZDRAZIL, Barbara, 2020. A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: case studies for rare diseases and COVID-19. *Journal of Cheminformatics*. Vol. 12, no. 1, p. 71. ISSN 1758-2946. Available from DOI: 10.1186/s13321-020-00474-z.

TUKEY, John Wilder, 1977. *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub. Co. Addison-Wesley series in behavioral science. ISBN 978-0-201-07616-5.

UMSCHEID, Craig A.; MARGOLIS, David J.; GROSSMAN, Craig E., 2011. Key Concepts of Clinical Trials: A Narrative Review. *Postgraduate Medicine*. Vol. 123, no. 5, pp. 194–204. ISSN 0032-5481, ISSN 1941-9260. Available from DOI: 10.3810/pgm.2011.09.2475.

VAMATHEVAN, Jessica; CLARK, Dominic; CZODROWSKI, Paul; DUNHAM, Ian; FERRAN, Edgardo; LEE, George; LI, Bin; MADABHUSHI, Anant; SHAH, Parantu; SPITZER, Michaela; ZHAO, Shanrong, 2019. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*. Vol. 18, no. 6, pp. 463–477. ISSN 1474-1776, ISSN 1474-1784. Available from DOI: 10.1038/s41573-019-0024-5.

VARSOU, Dimitra-Danai; NIKOLAKOPOULOS, Spyridon; TSOUMANIS, Andreas; MELAGRAKI, Georgia; AFANTITIS, Antreas, 2018. Enalos+ KNIME Nodes: New Cheminformatics Tools for Drug Discovery. In: MAVROMOUSTAKOS, Thomas; KELLICI, Tahsin F. (eds.). *Rational Drug Design*. New York, NY: Springer New York. Vol. 1824, pp. 113–138. ISBN 978-1-4939-8629-3. Available from DOI: 10.1007/978-1-4939-8630-9_7.

VENKATRAMAN, Vishwesh; PÉREZ-NUENO, Violeta I.; MAVRIDIS, Lazaros; RITCHIE, David W., 2010. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *Journal of Chemical Information and Modeling*. Vol. 50, no. 12, pp. 2079–2093. ISSN 1549-9596, ISSN 1549-960X. Available from DOI: 10.1021/ci100263p.

WANG, Fei; ZHANG, Ping; CAO, Nan; HU, Jianying; SORRENTINO, Robert, 2014. Exploring the associations between drug side-effects and therapeutic indications. *Journal of Biomedical Informatics*. Vol. 51, pp. 15–23. ISSN 15320464. Available from DOI: 10.1016/j.jbi.2014.03.014.

WERMUTH, J.; GANELLIN, C. R.; LINDBERG, P.; MITSCHER, L. A., 1998. Glossary for chemists of terms used in toxicology (IUPAC Recommendations 1993). *Pure and applied chemistry*. Vol. 70, no. 5, pp. 1129–1143.

WISHART, D. S., 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*. Vol. 34, no. 90001, pp. D668–D672. ISSN 0305-1048, ISSN 1362-4962. Available from DOI: 10.1093/nar/gkj067.

WONG, Chi Heem; SIAH, Kien Wei; LO, Andrew W, 2019. Estimation of clinical trial success rates and related parameters. *Biostatistics*. Vol. 20, no. 2, pp. 273–286. ISSN 1465-4644, ISSN 1468-4357. Available from DOI: 10.1093/biostatistics/kxx069.

WONG, Mei; MCALLISTER, Mark, 2017. Lead Identification/Optimization. In: KWONG, Elizabeth (ed.). *Oral Formulation Roadmap from Early Drug Discovery to Development*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 9–37. ISBN 978-1-118-90789-4. Available from DOI: 10.1002/9781118907894.ch2.

WOUTERS, Olivier J.; MCKEE, Martin; LUYTEN, Jeroen, 2020. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*. Vol. 323, no. 9, p. 844. ISSN 0098-7484. Available from DOI: 10.1001/jama.2020.1166.

XIE, Li; LI, Jerry; XIE, Lei; BOURNE, Philip E., 2009. Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network To Explain the Side Effects of CETP Inhibitors. *PLoS Computational Biology*. Vol. 5, no. 5, e1000387. ISSN 1553-7358. Available from DOI: 10.1371/journal.pcbi.1000387.

YANG, Xin; WANG, Yifei; BYRNE, Ryan; SCHNEIDER, Gisbert; YANG, Shengyong, 2019. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews*. Vol. 119, no. 18, pp. 10520–10594. ISSN 0009-2665, ISSN 1520-6890. Available from DOI: 10.1021/acs.chemrev.8b00728.

YIN, Yongmin; XU, Congying; GU, Shikai; LI, Weihua; LIU, Guixia; TANG, Yun, 2015. Quantitative Regression Models for the Prediction of Chemical Properties by an Efficient Workflow. *Molecular informatics*. Vol. 34, no. 10, pp. 679–688.

ZHANG, Ping; WANG, Fei; HU, Jianying; SORRENTINO, Robert, 2013. Exploring the relationship between drug side-effects and therapeutic indications. *AMIA ... Annual Symposium proceedings. AMIA Symposium*. Vol. 2013, pp. 1568–1577. ISSN 1942-597X.

List of abbreviations

ADME	Absorption, Distribution, Metabolism, Excretion
ASCII	American Standard Code for Information Interchange
ATC	Anatomical Therapeutic Chemical
AUC	Area Under the Curve
CIOMS	Council for International Organization of Medical Sciences
CSV	Comma Separated Values
DBM	Difference Between Medians
FCFP	Functional Class Fingerprint
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
InChI	IUPAC International Chemical Identifier
IUPAC	International Union of Pure and Applied Chemistry
KNIME	Konstanz Information Miner
MACCS	Molecular ACCess System
MedDRA	Medical Dictionary for Regulatory Activities
OVS	Overall Visible Spread
PAINS	Pan Assay Interference Structures
PDB	Protein Data Bank
PK	Pharmacokinetics
PMML	Predictive Model Markup Language
PR	Precision-Recall (curve)
QSAR	Quantitative Structure-Activity Relationship
ROC	Receiver Operating Characteristics (curve)
SE	Side Effect
SIDER	Side Effect Resource
SMARTS	SMiles ARbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
STITCH	Search Tool for Interacting Chemicals
TN	True Negative
TP	True Positive
UMLS	Unified Medical Language System

List of appendices

A List of additional files

Tab. A.1: Additional files

File name	Description
DISSERTATION_PROJECT.knar	main workflow file
filtered_drugs_dataset	dataset of drugs filtered from DrugBank
filtered_side_effects_dataset	dataset of side effects filtered from SIDER
filtered_indications_dataset	dataset of indications filtered from SIDER
filtered_targets_dataset	dataset of targets filtered from DrugBank
filtered_interacting_drugs_dataset	dataset of interacting drugs filtered from DrugBank
10_analyzed_side_effects	dataset of 10 analyzed side effects

Tab. A.2: Summary of filtered drug datasets

Dataset	Columns
filtered_drugs_dataset	drugbank_id, type, name, group, SMILES, atc_code, code_4
filtered_side_effects_dataset	STITCH_id_flat, UMLS_id, MedDRA_info_3, ATC, drugbank_id, name, SMILES
filtered_indications_dataset	STITCH_id_flat, drugbank_id, name, UMLS_label_indication, method_of_detection, concept_name, concept_type_MedDRA, UMLS_id_MedDRA_indication, concept_name_MedDRA_indication
filtered_targets_dataset	drugbank_id, name, target_id, target_name
filtered_interacting_drugs_dataset	drugbank_id, name, drugbank_id_int, int_drug_name

Tab. A.3: Description of analyzed data

Data type	Description
drugbank_id	drug molecule ID
STITCH_id	drug molecule ID
SMILES	drug structure format
ATC	Anatomical Therapeutic Chemical code of drug
UMLS_id SE_id	Side effect ID
MedDRA_info_3	side effect name
UMLS_id_MedDRA_indication	drug indication ID
concept_name_MedDRA_indication	drug indication name

B Correlation analysis

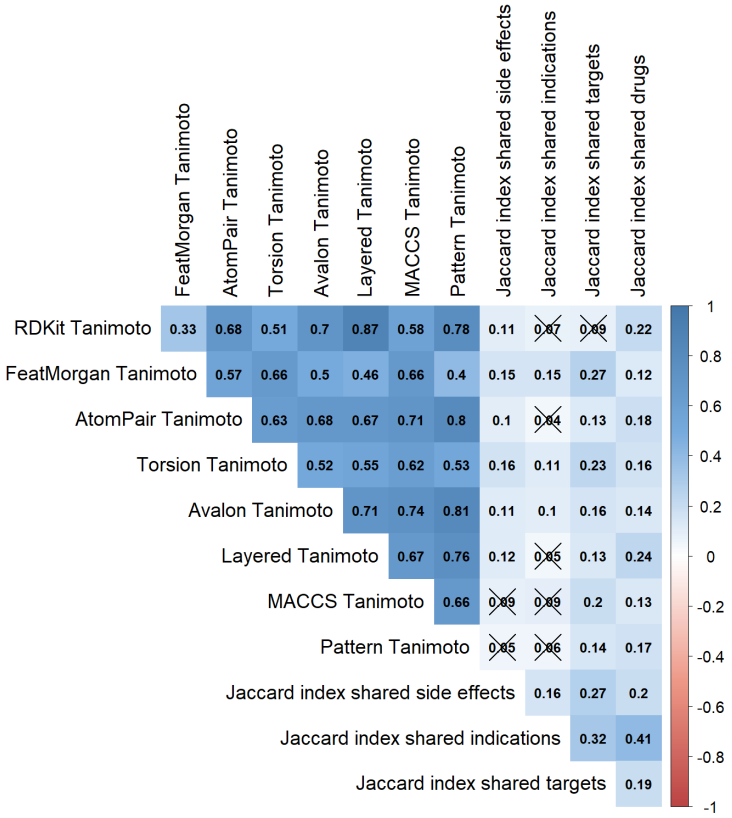


Fig. B.1: The exemplary color coded view of the feature correlation. Positive correlations are displayed in a blue color and negative correlations in a red color. The color intensity is proportional to the displayed correlation coefficients. Correlations with a p-value >0.05 are regarded as insignificant and the crosses are added.

Tab. B.1: The median values of the correlation distribution. Features with highest median correlation values.

Feature	**Median above 0.67 with following features**
RDKit	0.73 Avalon, 0.83 Layered
AtomPair	0.67 Layered, 0.67 Pattern
Avalon	0.76 Layered, 0.73 RDKit, 0.73 MACCS
Layered	0.83 RDKit, 0.76 Avalon, 0.67 AtomPair
MACCS	0.73 Avalon, 0.68 Layered
Pattern	0.67 AtomPair

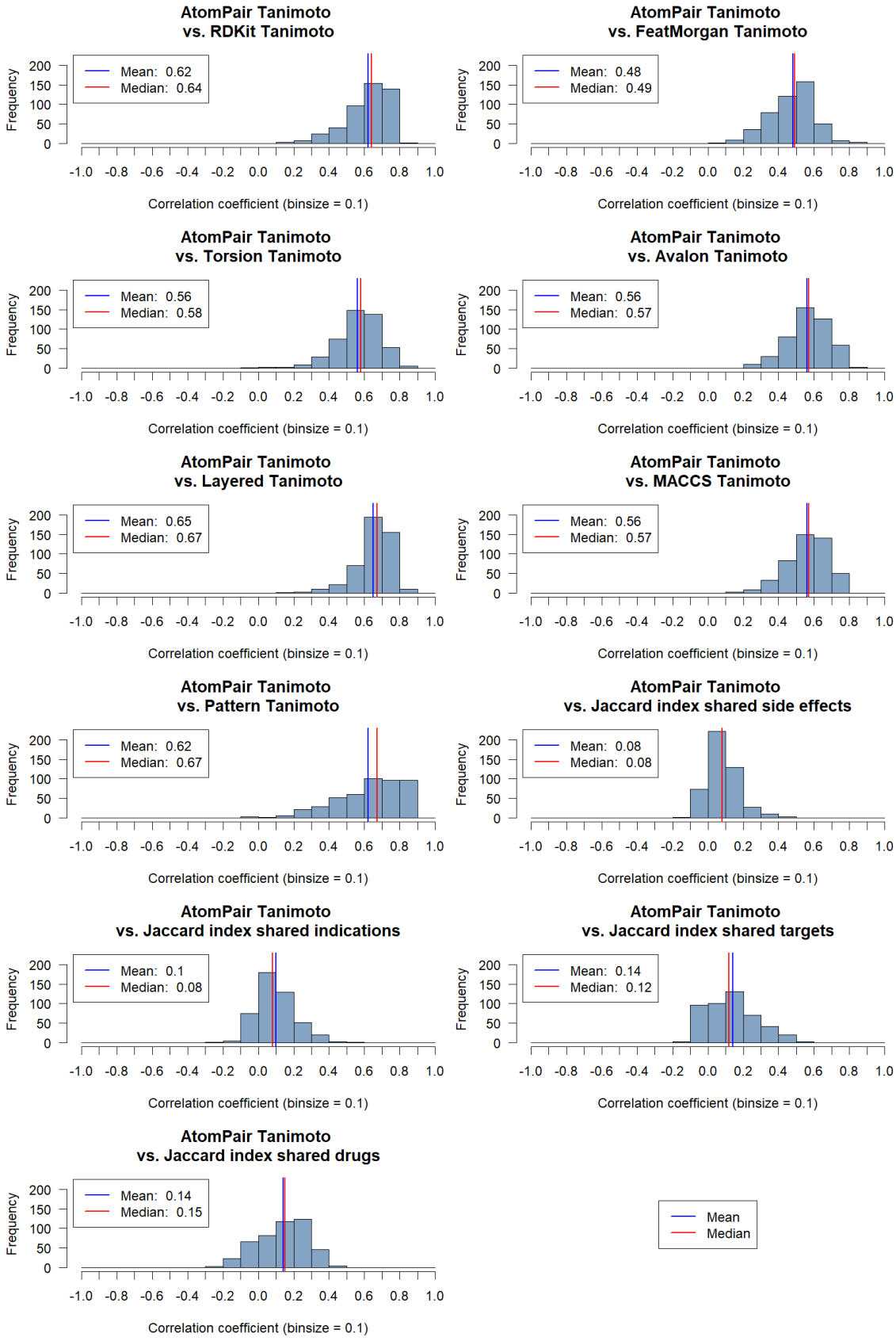


Fig. B.2: Correlation distribution of AtomPair Tanimoto similarity coefficient vs. similarity measures in all datasets

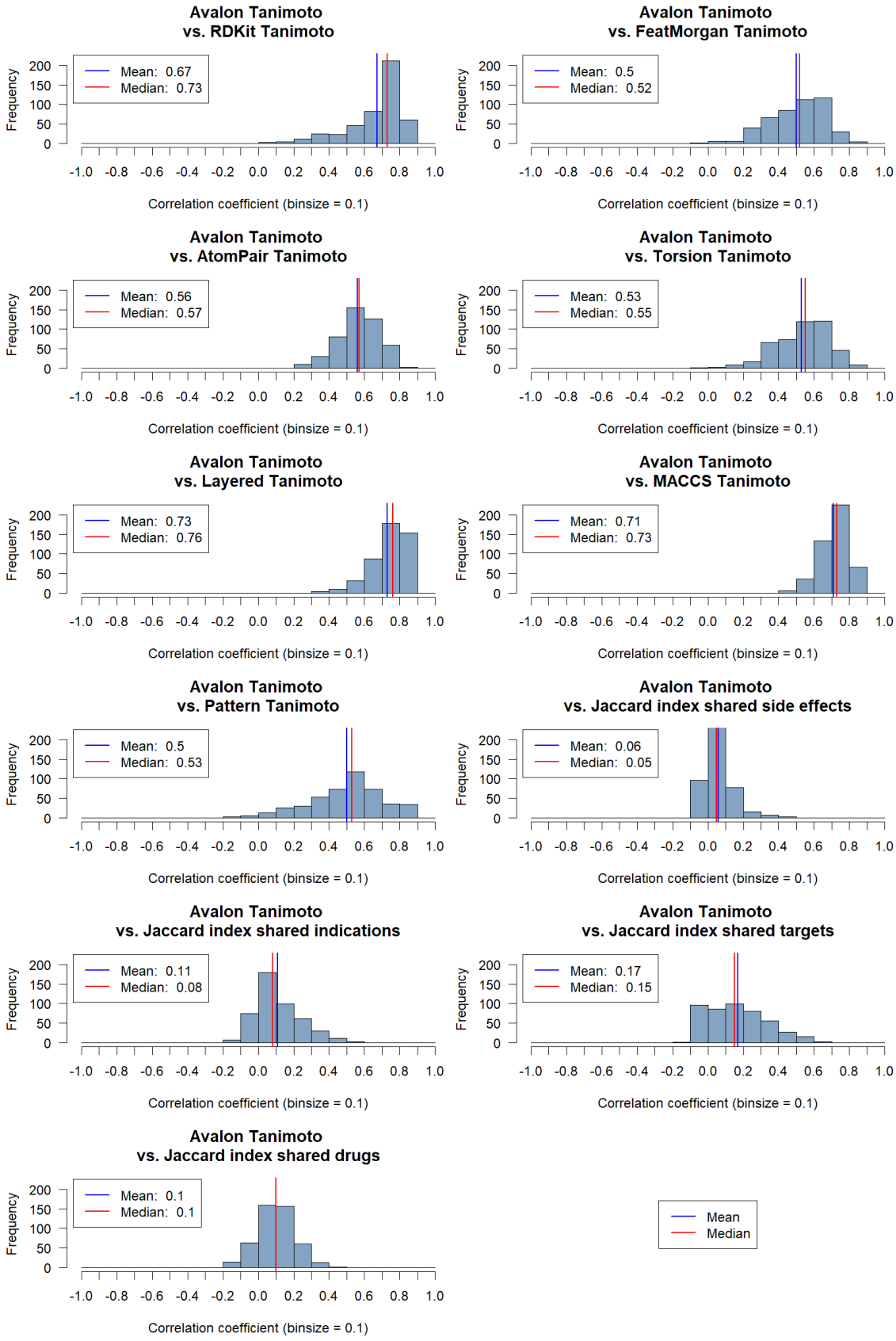


Fig. B.3: Correlation distribution of Avalon Tanimoto similarity coefficient vs. similarity measures in all datasets

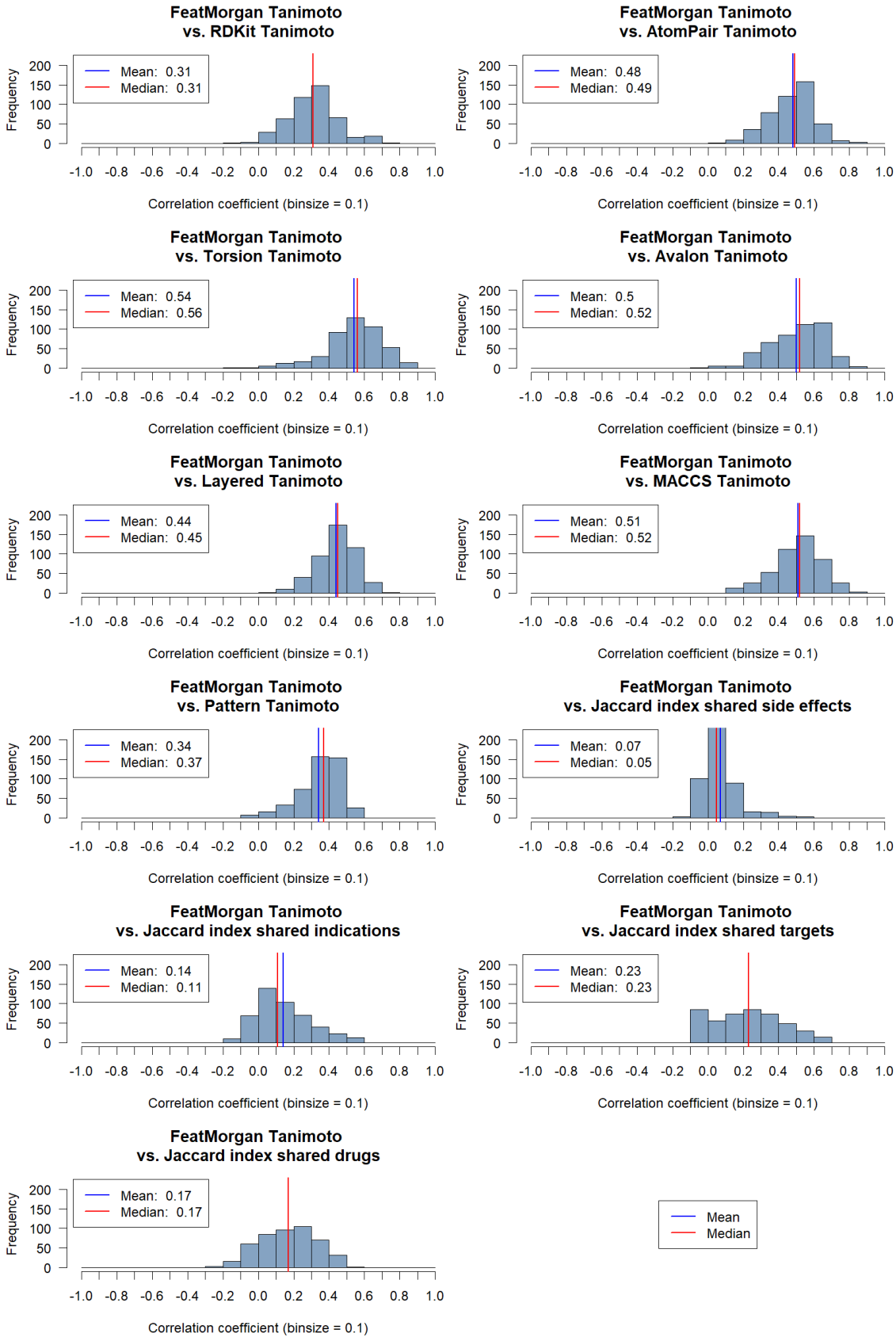


Fig. B.4: Correlation distribution of FeatMorgan Tanimoto similarity coefficient vs. similarity measures in all datasets

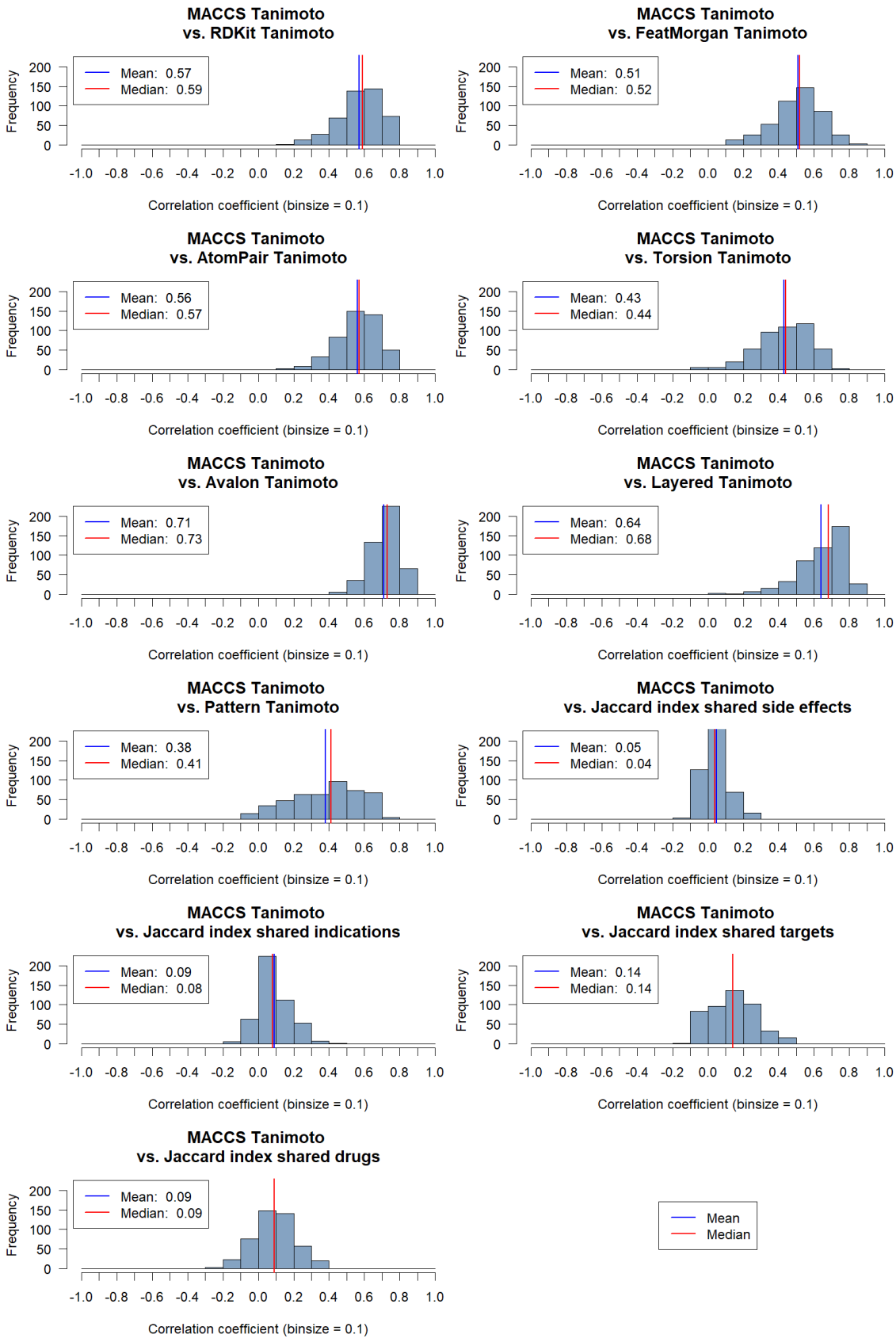


Fig. B.5: Correlation distribution of MACCS Tanimoto similarity coefficient vs. similarity measures in all datasets

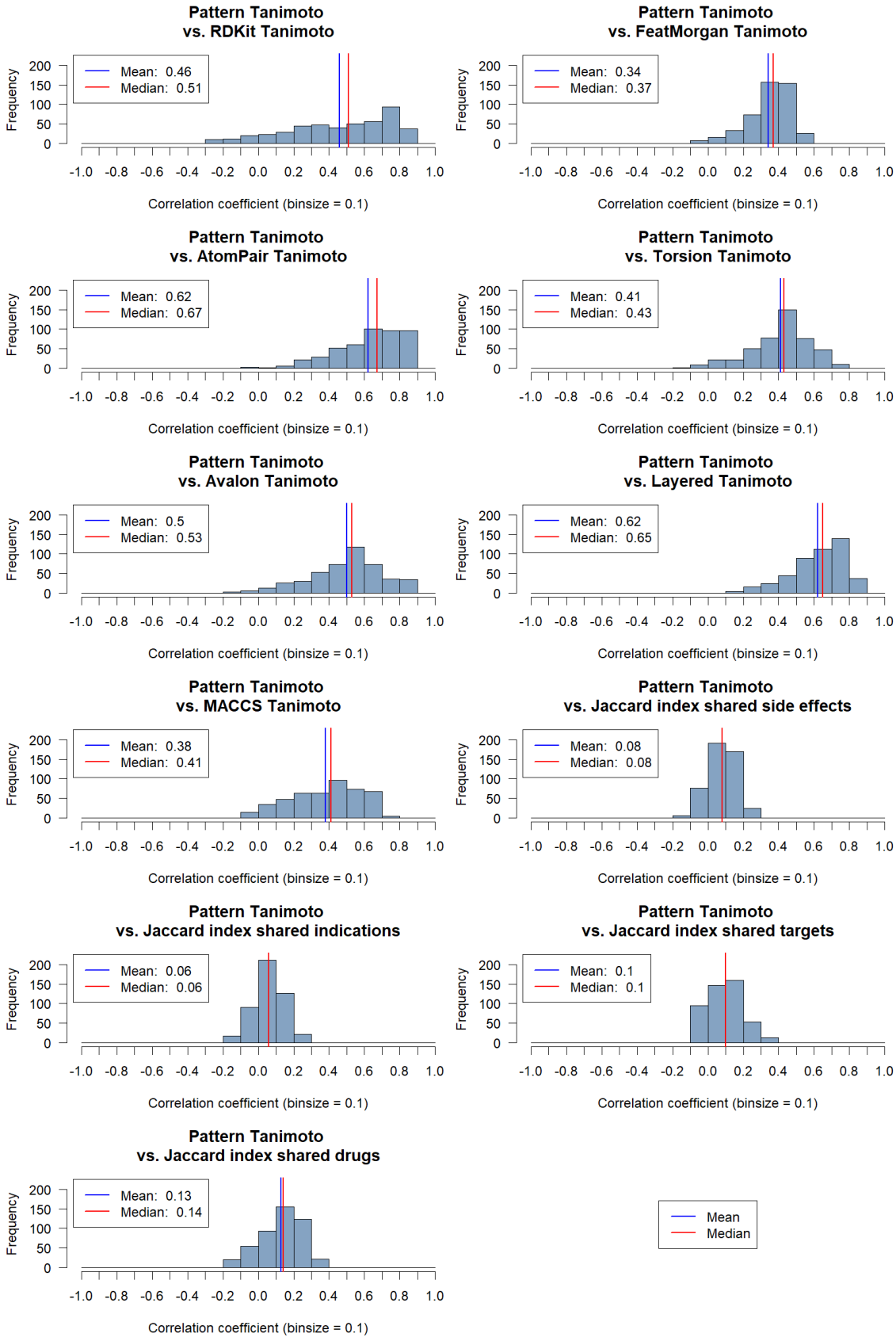


Fig. B.6: Correlation distribution of Pattern Tanimoto similarity coefficient vs. similarity measures in all datasets

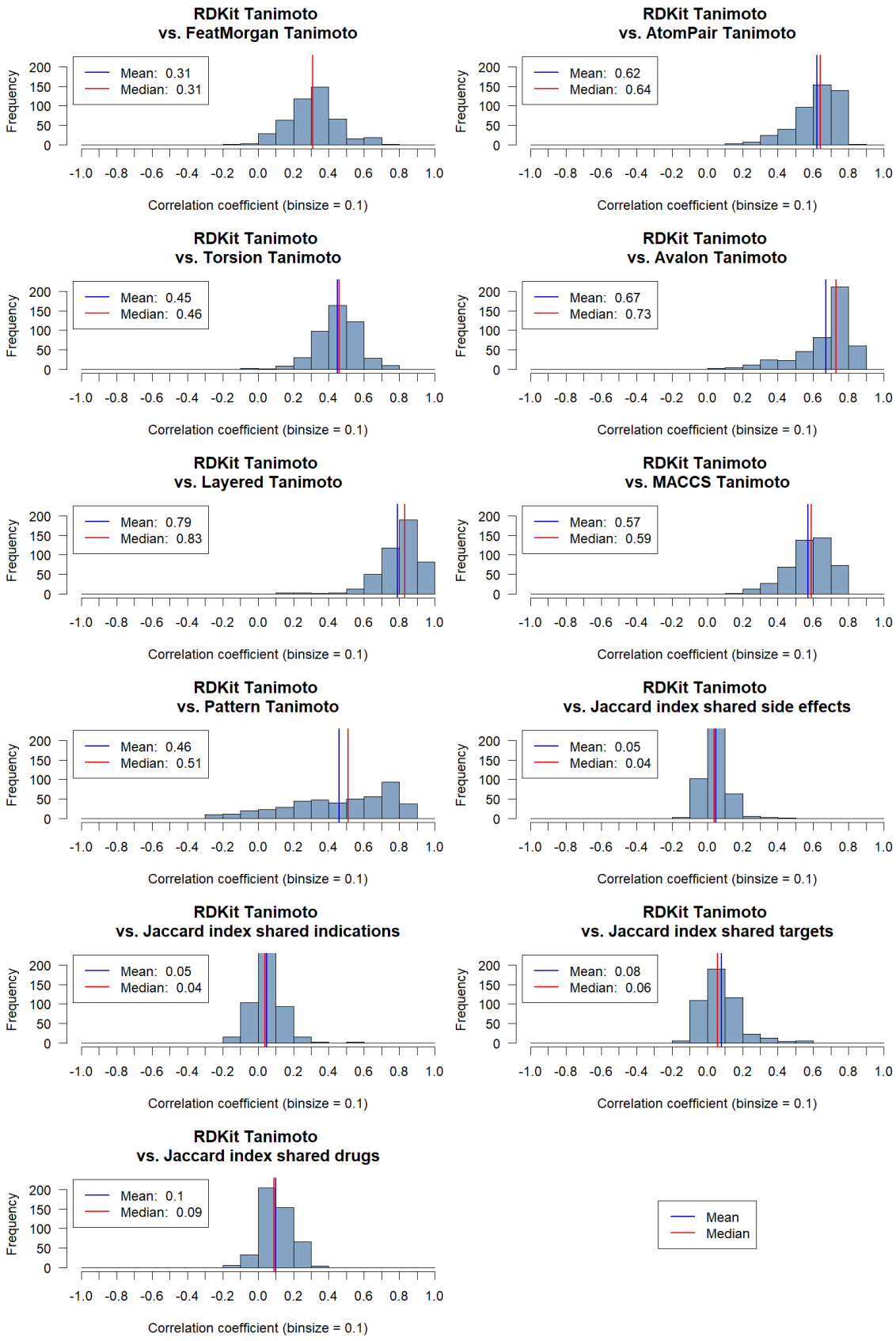


Fig. B.7: Correlation distribution of RDKit Tanimoto similarity coefficient vs. similarity measures in all datasets

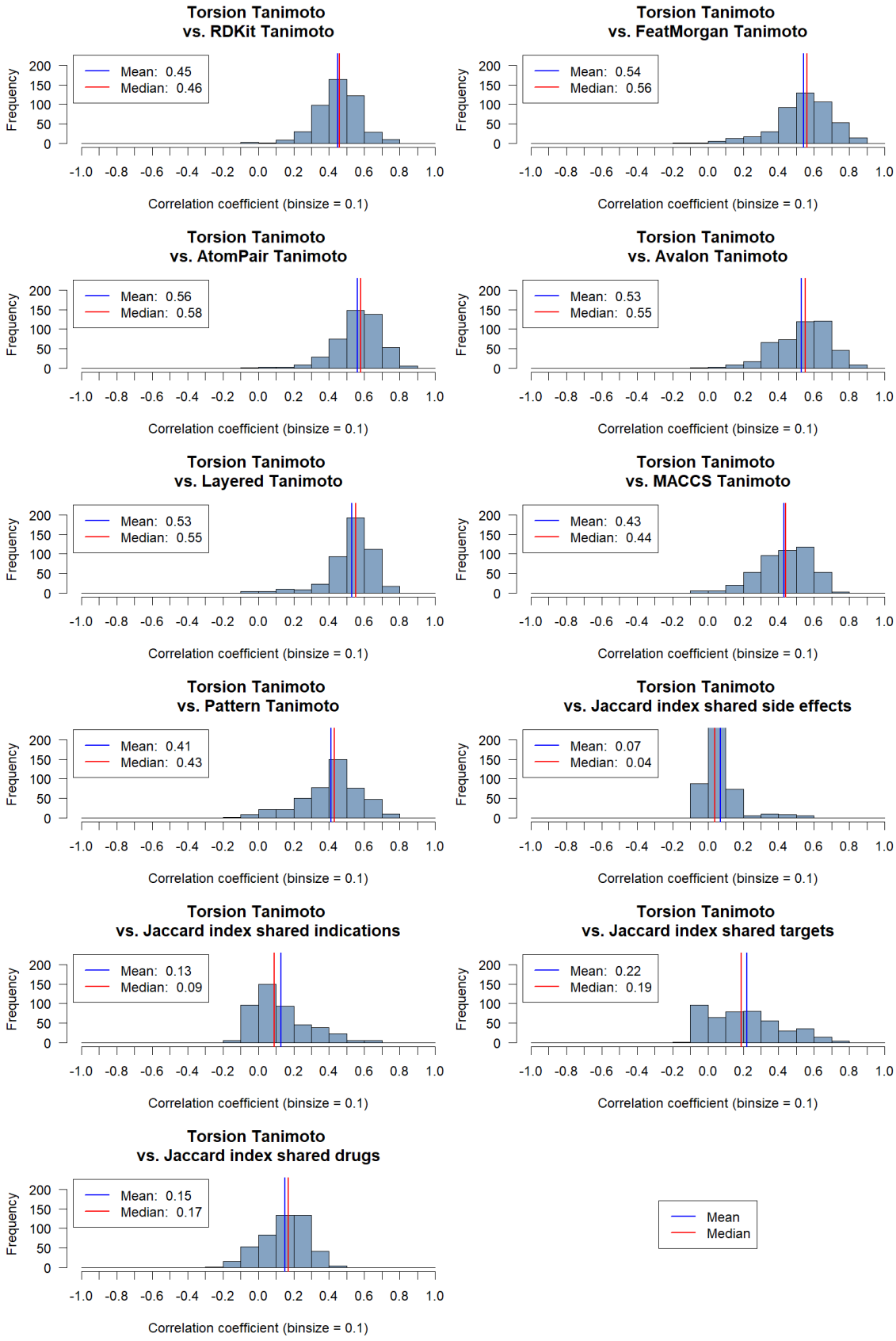


Fig. B.8: Correlation distribution of Torsion Tanimoto similarity coefficient vs. similarity measures in all datasets

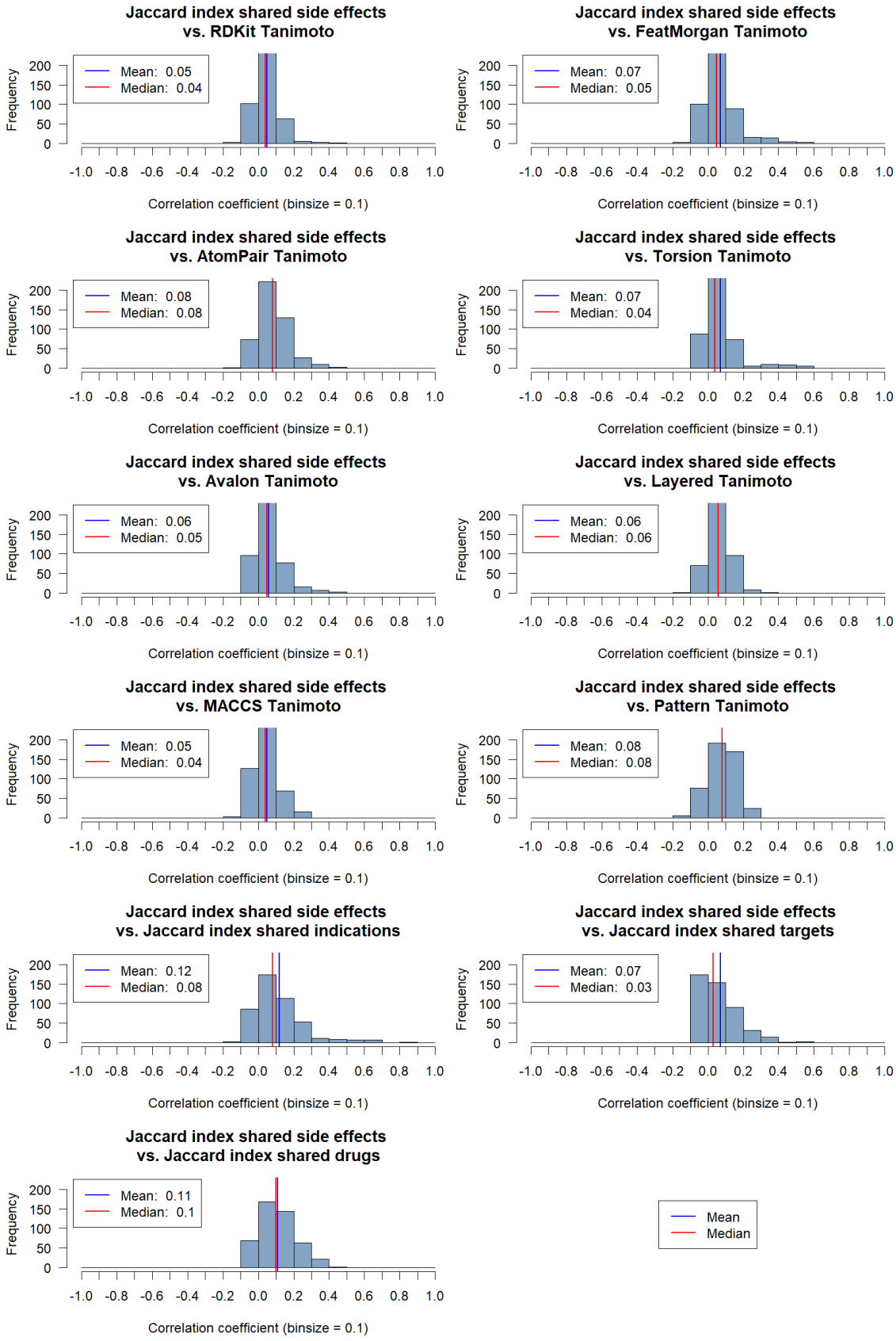


Fig. B.9: Correlation distribution of Jaccard index shared side effects vs. similarity measures in all datasets

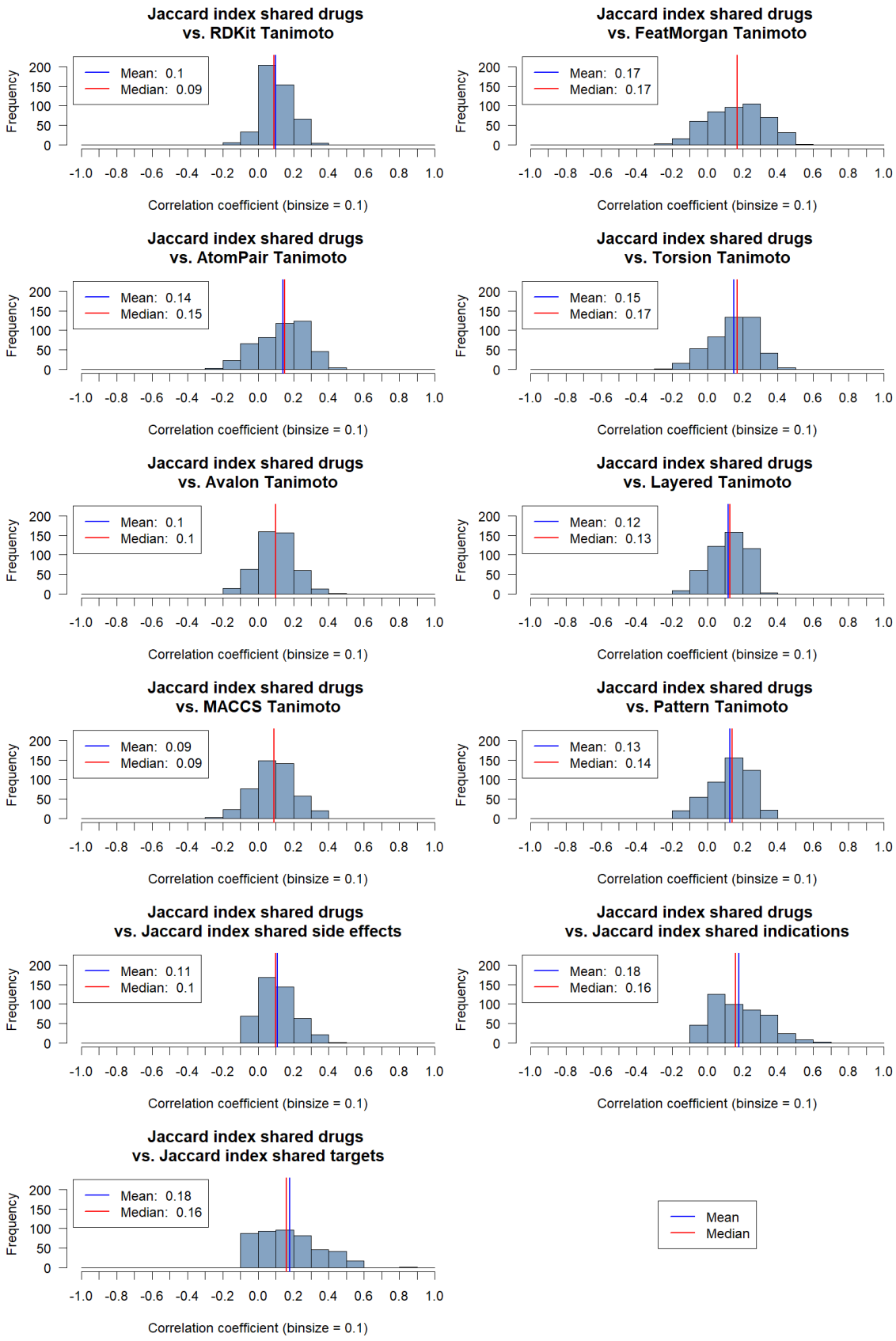


Fig. B.10: Correlation distribution of Jaccard index shared drugs vs. similarity measures in all datasets

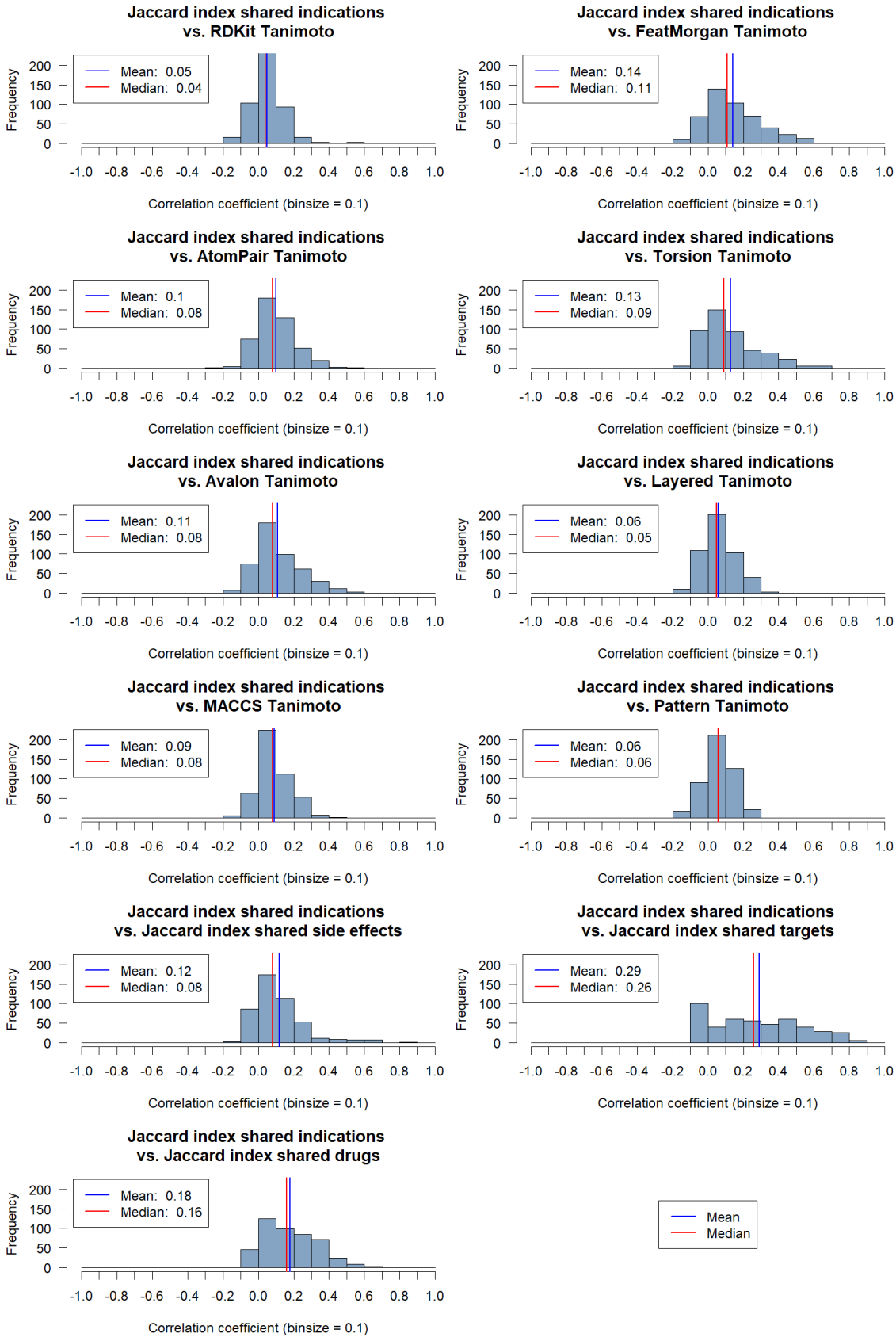


Fig. B.11: Correlation distribution of Jaccard index shared indications vs. similarity measures in all datasets

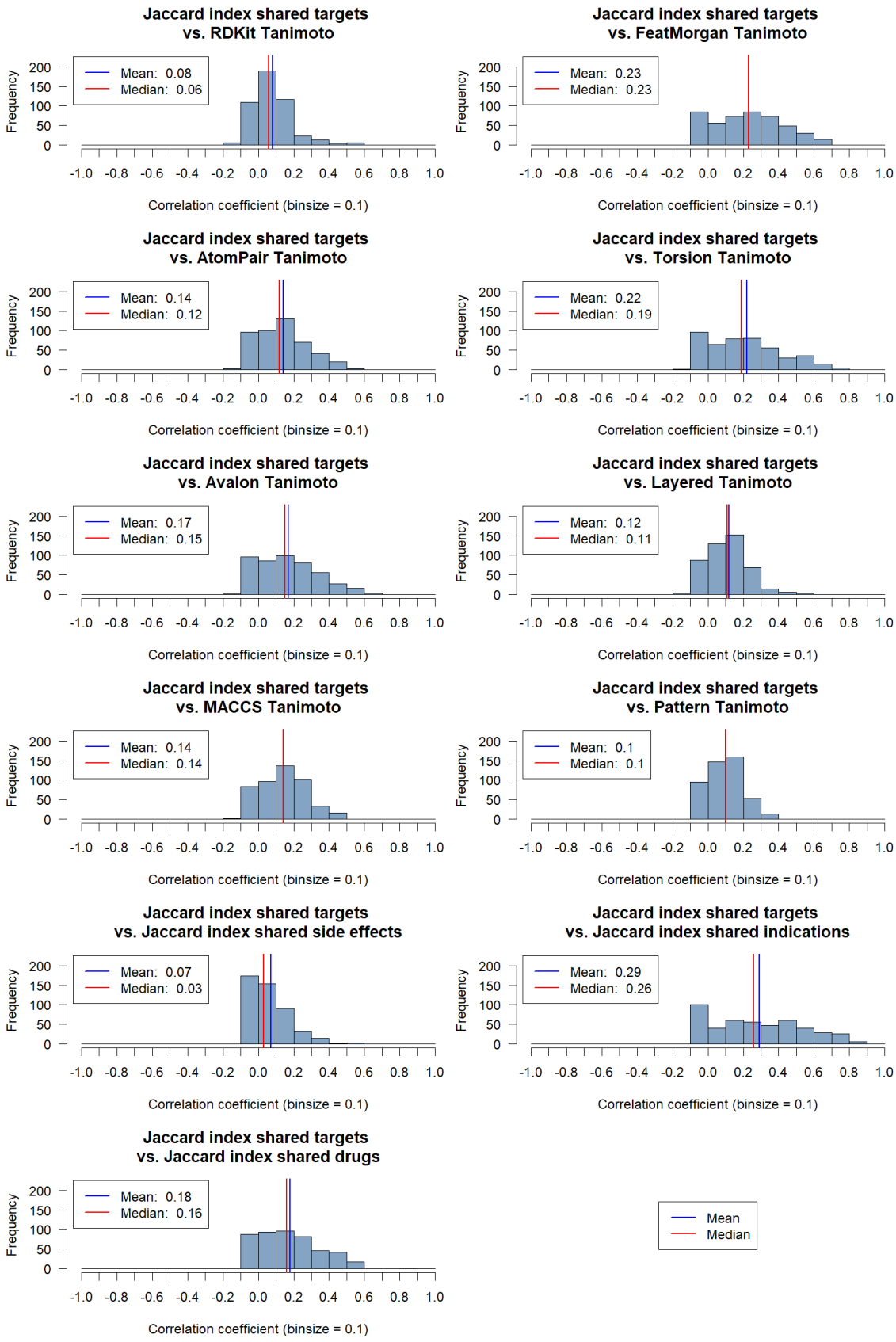


Fig. B.12: Correlation distribution of Jaccard index shared targets vs. similarity measures in all datasets

C Author's vita and list of publications

Pavlína Cicková (Koščová) 175408@vut.cz

EDUCATION

2009–2012	**Bachelor's degree in Molecular Biology and Genetics** Masaryk University in Brno, Faculty of Science Programme: Experimental Biology
2012–2014	**Master's degree in Chemoinformatics and Bioinformatics** Masaryk University in Brno, Faculty of Science Programme: Biochemistry
2012–2017	**Bachelor's degree in Business Management** Masaryk University in Brno, Faculty of Economics and Administration Programme: Economy and Management
2014–present	**Doctoral study** Brno University of Technology, Department of Biomedical Engineering Programme: Biomedical Technology and Bioinformatics

SELECTED TRAININGS DURING DOCTORAL STUDY

2014	**Kurz základů vědecké práce**, Akademie věd České republiky
2015	**CCG MOE Workshop**, University of Hamburg
2015	**IEEE Xplore Trainings about Authorship in Czech Republic**, Brno University of Technology
2015	**EUROPIN Summer School on Drug Design**, University of Vienna
2016	**Medicinal Chemistry: The Molecular Basis of Drug**, DiscoveryX, online
2016	**Sense and Sensibility Visual Design Principles for Scientic Data**, CEITEC, Masaryk University
2017	**Workshops at 44th EMWA Conference**, ICC, Birmingham
2017	**Letní škola matematické biologie**, Telč, Masaryk University
2017	**NGS Analytical Pipeline Design and Management**, CEITEC
2018	**3rd Advanced *in silico* Drug Design workshop/hackathon**, Palacky University Olomouc
2018	**Medical and Science Writing Skills for Academia and Industry**, StGilesMedical, Berlin
2018	**Erasmus+ Internship**, StGilesMedical, Berlin

TEACHING EXPERIENCE

at Brno University of Technology

2014/15–2017/18	**FMOL: Molecular biology** – practicals tutor
2014/15–2017/18	**FSYS: Systems biology** – practicals tutor
2015/16–2016/17	**Bachelor theses supervisor**
2015	Tutor at training workshops for masters and doctoral students ***How to find the right drug?***

AWARDS

2017	**Award for the Best Educational Scientific Short Film** Neuron Prima ZOOM, Prague
2017	**Talented PhD Student Award at Faculty of Electrical Engineering and Communication**, Brno University of Technology
2018	**Second Place in *in silico* Drug Design Challenge** Advanced *in silico* Drug Design workshop/hackathon 2018 Palacky University Olomouc

IDENTIFIERS AND SCIENTIFIC IMPACT

current state in December 2021

ORCID	0000-0003-1341-7975
Web of Science ID	AAS-4760-2021
Scopus ID	57190748346 (Koščová, P.) / 57214141647 (Cicková, P.)
Number of records	12
H-index	3
Total times cited	77

PROJECTS

2017–2018	**Relationship between butanol efflux and butanol tolerance of Clostridia**, Grant Agency of the Czech Republic GA17-00551S – fellow researcher

JOURNAL ARTICLES

KOŠČOVÁ, P.; PROVAZNÍK, I., 2016. Pharmacophore modelling used in rational drug design. *Chemické listy.* Vol. 110, no. 8, pp. 575-580. ISSN 0009-2770.

PATÁKOVÁ, P.; KOLEK, J.; SEDLÁŘ, K.; **KOŠČOVÁ, P.**; BRANSKÁ, B.; KUPKOVÁ, K.; PAULOVÁ, L.; PROVAZNÍK, I., 2017. Comparative analysis of high butanol tolerance and production in clostridia. *Biotechnology Advances.* Vol. 35, no. 8, pp. 1-38. ISSN: 0734-9750.

SEDLÁŘ, K.; **KOŠČOVÁ, P.**; VASYLKIVSKA, M.; BRANSKÁ, B.; KOLEK, J.; KUPKOVÁ, K.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2018. Transcription profiling of butanol producer Clostridium beijerinckii NRRL B-598 using RNA-Seq. *BMC Genomics.* Vol. 19, no. 415, s. 1 (1 s.). ISSN: 1471-2164.

MADĚRÁNKOVÁ, D.; MIKALOVÁ, L.; STROUHAL, M.; VADJÁK, Š.; KUKLOVÁ, I.; POSPÍŠILOVÁ, P.; KRBKOVÁ, L.; **KOŠČOVÁ, P.**; PROVAZNÍK, I.; ŠMAJS, D., 2019. Identification of positively selected and recombinant genes in uncultivable pathogenic treponemes: syphilis-, yaws-, and bejel-causing strains differ in sets of genes showing adaptive evolution. *PLoS Neglected Tropical Diseases.* Vol. 13, no. 6. DOI: 10.1371/journal.pntd.0007463

PATÁKOVÁ, P.; BRANSKÁ, B.; SEDLÁŘ, K.; VASYLKIVSKA, M.; JUREČKOVÁ, K.; KOLEK, J.; **KOŠČOVÁ, P.**; PROVAZNÍK, I., 2019. Acidogenesis, solventogenesis, metabolic stress response and life cycle changes in Clostridium beijerinckii NRRL B-598 at the transcriptomic level. *Scientific Reports.* Vol. 9, no. 1, pp. 1-21. ISSN: 2045-2322.

WALKER, S.; MALLADI, A.; **CICKOVA, P.**; DAVIES, R.; O´CONNOR, H., 2019. An introduction to medical affairs for medical writers. *Medical Writing.* Vol. 28, no. 4, pp. 39-43. ISSN: 20474814

POSTERS

KOŠČOVÁ, P.; ROY, S.; PROVAZNÍK, I., 2016. Use of 3D QSAR Pharmacophore Modelling in the Research of Influenza. In: *Proceedings of the 4th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 1-4. ISBN: 978-80-86238-91-3.

KOŠČOVÁ, P.; SEDLÁŘ, K.; KUPKOVÁ, K.; KOLEK, J.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2017. Tertiary Structure Prediction of Potential Efflux-Pump Protein in Clostridium beijerinckii NRRL B-598. In: *Proceedings of the 5th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 26-30. ISBN: 978-80-86238-62-3.

JUREČKOVÁ, K.; **KOŠČOVÁ, P.**; SEDLÁŘ, K.; KOLEK, J.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2018. *In Silico* Prediction of Genes Coding Efflux Pumps in Clostridium Beijerinckii NRRL B-598. In: *Proceedings of the 6th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 1-5. ISBN: 978-80-86238-83-8.

CONFERENCE PROCEEDINGS

KOŠČOVÁ, P.; PROVAZNÍK, I, 2016. Comparative Analysis of Common and Unique Targets in Drug Resistant Strains of Borrelia burgdorferi. In: *Proceedings of the 22nd Conference STUDENT EEICT 2016.* Brno: Brno University of Technology, pp. 523-527. ISBN: 978-80-214-5350-0.

ČMIEL, V.; SVOBODA, O.; **KOŠČOVÁ, P.**; PROVAZNÍK, I., 2016. Smartphone based point-of-care detector of urine albumin. In: *Proceedings of SPIE Volume 9715. Optical Diagnostics and Sensing XVI: Toward Point-of-Care Diagnostics.* San Francisco: SPIE - International Society for Optics and Photonics, pp. 99-103. ISBN: 9781628419498.

KOŠČOVÁ, P.; ROY, S.; PROVAZNÍK, I., 2016. Use of 3D QSAR Pharmacophore Modelling in the Research of Influenza. In: *Proceedings of the 4th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 1-4. ISBN: 978-80-86238-91-3.

SEDLÁŘ, K.; **KOŠČOVÁ, P.**; KUPKOVÁ, K.; VASYLKIVSKA, M.; BRANSKÁ, B.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2017. Genome mining for biorefinery use. In: *Biotech 2017 and 7th czech-swiss symposium with Exhibition Book of Abstracts.* Prague: pp. 48-48. ISBN: 978-80-7080-989-1.

SEDLÁŘ, K.; BRANSKÁ, B.; KUPKOVÁ, K.; **KOŠČOVÁ, P.**; KOLEK, J.; VASYLKIVSKA, M.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2017. Analysis of Clostridium beijerinckii NRRL B-598 coding regions using RNA-Seq data of a closely related strain. In: *Proceedings of the 5th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 87-91. ISBN: 978-80-86238-62-3.

KOŠČOVÁ, P.; SEDLÁŘ, K.; KUPKOVÁ, K.; KOLEK, J.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2017. Tertiary Structure Prediction of Potential Efflux-Pump Protein in Clostridium beijerinckii NRRL B-598. In: *Proceedings of the 5th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 26-30. ISBN: 978-80-86238-62-3.

KOŠČOVÁ, P., 2017. Molecular docking study of potential drug candidates against borreliosis. In: *Proceedings of the 23rd Conference STUDENT EEICT 2017.* Brno: Brno University of Technology, pp. 339-343. ISBN: 978-80-214-5496- 5.

MUSILOVÁ, J.; **KOŠČOVÁ, P.**, 2018. Network Analysis of Proteins Associated with Schizophrenia. In: *Proceedings of the 24rd Conference STUDENT EEICT 2018.* Brno: Brno University of Technology, pp. 273-275. ISBN: 978-80-214-5614-3.

JUREČKOVÁ, K.; **KOŠČOVÁ, P.**; SEDLÁŘ, K.; KOLEK, J.; PATÁKOVÁ, P.; PROVAZNÍK, I., 2018. *In Silico* Prediction of Genes Coding Efflux Pumps in Clostridium Beijerinckii NRRL B-598. In: *Proceedings of the 6th International Conference on Chemical Technology.* Prague: Czech Society of Industrial Chemistry, pp. 1-5. ISBN: 978-80-86238-83-8.