Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

MSc. Yingying Jiang

born in Shandong, China

Oral-examination:

Systems and chemical biology approaches to study cell function and response to toxins

Referees: Prof. Dr. Rob Russell

Prof. Dr. Stefan Wölfl

CONTRIBUTIONS

The chapter III of this thesis was submitted for publishing under the title “Drug mechanism predominates over toxicity mechanisms in drug induced gene expression” by Yingying Jiang, Tobias C. Fuchs, Kristina Erdeljan, Bojana Lazerevic, Philip Hewitt, Gordana Apic & Robert B. Russell. For chapter III, text phrases, selected tables, figures are based on this submitted manuscript that has been originally written by myself.

ABSTRACT

Toxicity is one of the main causes of failure during drug discovery, and of withdrawal once drugs reached the market. Prediction of potential toxicities in the early stage of drug development has thus become of great interest to reduce such costly failures. Since toxicity results from chemical perturbation of biological systems, we combined biological and chemical strategies to help understand and ultimately predict drug toxicities.

First, we proposed a systematic strategy to predict and understand the mechanistic interpretation of drug toxicities based on chemical fragments. Fragments frequently found in chemicals with certain toxicities were defined as structural alerts for use in prediction. Some of the predictions were supported with mechanistic interpretation by integrating fragment-chemical, chemical-protein, protein-protein interactions and gene expression data.

Next, we systematically deciphered the mechanisms of drug actions and toxicities by analyzing the associations of drugs' chemical features, biological features and their gene expression profiles from the TG-GATEs database. We found that *in vivo* (rat liver) and *in vitro* (rat hepatocyte) gene expression patterns were poorly overlapped and gene expression responses in different species (rat and human) and different tissues (liver and kidney) varied widely.

Eventually, for further understanding of individual differences in drug responses, we reviewed how genetic polymorphisms influence the individual's susceptibility to drug toxicity by deriving chemical-protein interactions and SNP variations from Mechismo database. Such a study is also essential for personalized medicine.

Overall, this study showed that, integrating chemical and biological in addition to genetic data can help assess and predict drug toxicity at system and population levels.

Keywords: mechanism of drug-action, toxicogenomics, biological features, chemical features, 1000 Genomes

ZUSAMMENFASSUNG

Toxizitäten von Arzneimitteln sind eine der Hauptursachen des Scheiterns während des Wirkstoff-Entdeckungsprozesses - und der Vom-Markt-Nahme, sollten diese bereits den Markt erreicht haben. Daher ist die Vorhersage potentieller Toxizitäten im frühen Stadium der Arzneimittelentwicklung von großem Interesse geworden, um diesen langen und teuren Prozess zu verbessern. Da Toxizitäten aus der Störung biologischer Systeme durch Chemikalien resultieren, kombinierten wir biologische und chemische Strategien, um vollständige Vorhersagemodelle für Arzneimitteltoxizitäten bereitzustellen.

Zuerst haben wir eine systematische Strategie vorgeschlagen, um die mechanistische Interpretation von Wirkstofftoxizitäten auf der Basis chemischer Fragmente vorherzusagen und zu verstehen. Fragmente, die in hohem Maße mit bestimmten Toxizitäten zusammenhängen, wurden als Strukturalarmhinweise für die Verwendung in der Toxizitätsvorhersage definiert. Einige der Vorhersagen wurden mittels mechanistischer Interpretation durch Integration von Fragment-Chemie-, Chemie-Protein-, Protein-Protein-Interaktionen und Genexpressionsdaten unterstützt.

Als nächstes haben wir systematisch die Mechanismen von Arzneimittelwirkungen und -toxizitäten durch Analyse der Assoziationen zwischen chemischen Merkmalen und biologischen Merkmalen von Arzneimitteln und ihren Genexpressionsprofilen aus der TG-GATE-Datenbank entschlüsselt. In der Zwischenzeit fanden wir, dass sich *in vivo* (Rattenleber) und *in vitro* (Rattenhepatozyten) Genexpressionsmuster nur selten ähneln und -reaktionen bei verschiedenen Spezies (Ratte und Mensch), verschiedenen Geweben (Leber und Niere) weit variieren.

Zum besseren Verständnis der individuellen Unterschiede in den Medikamentenreaktionen untersuchten wir, wie genetische Polymorphismen die Anfälligkeit des Individuums gegenüber der Wirkstofftoxizität beeinflussen, indem wir chemische Protein-Interaktionen und SNP-Variationen aus der Mechismo-Datenbank abgeleitet haben. Eine solche Studie ist auch wichtig für die personalisierte Medizin.

Insgesamt zeigte diese Studie, dass die Integration von chemischen Merkmalen (z. B. chemische Fragmente, chemische Struktur), biologische Merkmale - einschließlich Genexpression und genetische Daten - die Toxizität des Wirkstoffs auf System- und Populationsniveau einschätzen und voraussagen können.

Schlüsselwörter: Mechanismus der Arzneimittelwirkung, Toxikogenomik, biologische Merkmale, chemische Eigenschaften, 1000 Genome

ACKNOWLEDGMENTS

This thesis is a milestone in my four years of work at Heidelberg University and specifically within the Professor Russell's laboratory. My experience in Heidelberg has been nothing short of amazing. Since high school, I started to dream of studying in Heidelberg. Until 2012, I have been given unique opportunities, starting as a PhD research assistant. Throughout these years in the lab, I have learned a lot from the whole team. Here, I would like to express my gratitude to all those who helped me during the writing of this thesis.

My deepest gratitude goes first and foremost to Professor Rob Russell, my supervisor, for his constant encouragement and guidance. He has walked me through all the stages of the writing of this thesis. I appreciate his continuous support and his promotion of my scientific work.

This thesis is also the result of many experiences I have encountered at Russell's group from dozens of remarkable individuals who I also wish to acknowledge, Gordana, Yvonne, Matthew, Qianhao, Oliver, Francesco, Juan-Carlos, Gurdeep and Chu-Harn.

Then, I would like to express my gratitude to Professor Wölfl, who gave me valuable advice on the thesis. I am also greatly indebted to the professors and co-workers at the Cell Networks, BioQuant, Biochemie Zentrum Heidelberg (BZH), Cambridge Cell Networks Ltd, Merck KgaA who have instructed and helped me a lot in the past four years.

I would also like to thank PhD programs in Institute of Pharmacy and Molecular Biotechnology and BZH for providing lectures, a great network of researchers, and academic training.

Finally, I want to say thanks to my beloved family for their loving considerations in China, who have been doing the utmost to support me. I sincere thank Norbert and Rodica for sharing their caring thoughts and taking care of me. I also owe my gratitude to all my friends who gave me their help and concern in helping me work out my problems during the difficult course of the living in Germany.

TABLE OF CONTENTS

Chapter Page

CONTRIBUTIONS i

ABSTRACT iii

ZUSAMMENFASSUNG v

ACKNOWLEDGMENTS vii

TABLE OF CONTENTS ix

ABBREVIATIONS xiii

CHAPTER I: Introduction 1

1.1 Cheminformatics-based predictions in toxicology 1

1.2 Bioinformatics-based predictions in toxicology 3

1.3 Integrated chemical and biological modeling-based predictions in toxicology 6

1.4 Predictive toxicology at human population level 8

1.5 Dissertation outline 10

CHAPTER II: Chemical fragments as foundations for understanding the toxic effects of chemicals on biological systems 13

2.1 Abstract 13

2.2 Introduction 13

2.3 Methods 17

2.3.1 Data preparation 17

2.3.2 Extracting toxicity information 18

2.3.3 Classifying fragments and dataset diversity 18

2.3.4 Toxic fragments analysis 19

2.3.5 Prediction methods 19

2.3.6 Performance evaluation 20

2.4 Results 21

2.4.1 Datasets 21

2.4.2 Toxic fragment analysis 21

2.4.3 Prediction via support vector machines 25

2.4.4 Prediction via sparse canonical correlation analysis 29

2.4.5 Mechanism study of toxicity 34

2.5 Conclusion and discussion 37

CHAPTER III: Deciphering mechanisms of drug action and toxicity by integrating gene expression signatures, biological features and chemical features 39

3.1 Abstract 39

3.2 Introduction 39

3.3 Methods 41

3.3.1 Toxicogenomics datasets 41

3.3.2 Microarray data and pathological data processing 41

3.3.3 Principal component analysis (PCA) and hierarchical clustering analysis (HCA)42

3.3.4 Biological data collection and visualization 42

3.3.5 Chemical data collection and visualization 42

3.3.6 Sparse canonical correlation analysis (SCCA) 43

3.4 Results 43

3.4.1 Gene expression quality controls 43

3.4.2 Gene expression profiles reflect common pharmacology 50

3.4.3 Comparison of *in vivo* and *in vitro* gene expression 56

3.4.4 Comparison between Rat & Human data 60

3.4.5 Liver and kidney gene expression comparison *in vivo* 61

3.4.6 Agreement between gene expression profiles and chemical features 61

3.4.7 Predicting pathology signatures based on chemical fragments and gene expression data 65

3.5 Conclusion and discussion 67

CHAPTER IV: Common, functional mutations in human subpopulations potentially causing individual susceptibility to toxicities or other pathologies 71

4.1 Abstract 71

4.2 Introduction 71

4.3 Methods 74

4.3.1 Processing Data from 1000G 74

4.3.2 Deriving protein altering variants information 74

4.3.3 Gene enrichment analysis 74

4.3.4 Relating chemical molecules to genes 75

4.3.5 Predicting the effects of missense variations 75

4.4 Results 75

4.4.1 Gene function enrichment 75

4.4.2 Variants at the chemical binding sites 79

4.4.3 Influences of variations on toxicity responses 83

4.5 Conclusion and discussion 86

CHAPTER V: General discussion and conclusions 89

5.1 Summary of major findings 89

5.1.1 Data-driven identification of structural alerts for predicting drug toxicity 89

5.1.2 Gene expression analysis provided mechanisms of drug toxicity 89

5.1.3 Genetic polymorphisms analysis provided mechanisms of individual differences of drug toxicity 90

5.2 Contributions and practical implications 90

5.2.1 Structural alert concept as applied in drug discovery 90

5.2.2 Integrative gene expression-chemical features for toxicity assessment 91

5.2.3 Pharmacogenomic characterization of drug response enables personalized medicine 91

5.3 Limitations and future solutions 92

5.3.1 Chemical data limitations and future solutions 92

5.3.2 Biological data limitations and future solutions 92

5.3.3 Methodological limitations and future solutions 92

5.4 Epilogue 93

References 95

Supporting Information 111

ABBREVIATIONS

ADME	Absorption, Distribution, Metabolism, and Excretion
AOPs	Adverse Outcome Pathways
AUC	Area under the ROC Curve
DILI	Drug-Induced Liver Injury
DNN	Deep Neural Network
EGFR	Epidermal Growth Factor Receptor
FP	False Positive
GC-MS	Gas Chromatography Mass Spectrometry
GWAS	Genome-Wide Association Study
HTS	High-Throughput Screening
kNN	k-Nearest Neighbors
LC-MS	Liquid Chromatography Mass Spectrometry
LD_{50}	Lethal Dose that kills 50% of the test animals in an experiment
MAF	Minor Allele Frequency
MeSH	Medical Subject Headings
MLR	Multiple Linear Regression
MoA	Mechanism of Action
MSG	Monosodium Glutamate
NMR	Nuclear Magnetic Resonance
Obs/exp	Ratio Observed/expected
PCA	Principal Component Analysis
PDB	Protein Data Bank
PLS	Partial Least Square
QSAR	Quantitative Structure-Activity Relationship
R&D	Research and Development
ROS	Reactive Oxygen Species
SAs	Structural Alerts
SCCA	Sparse Canonical Correlation Analysis
SMARTS	Smiles Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry Specification
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
TP	True Positive

CHAPTER I: Introduction

Despite overall dramatic increase in pharmaceutical R&D spending over the past few decades, relatively few drugs reached the market due to high failure rates at the clinical and pre-clinical testing stages (Kaitin, 2010). A major cause of these failures is toxicity (Paul et al., 2010). According to reports, only 8% of drug candidates in clinical trials entered the market, and around 20% of failures in the late drug discovery process were caused by the presence of toxicities (Muster et al., 2008). A small fraction of drugs were prohibited by regulatory agencies or withdrawn later after launching to the market due to the toxicity issues (e.g., adverse drug reactions, drug side effects) (Siramshetty et al., 2016).

Despite the clinical impacts of drug toxicity, methods for monitoring them remain limited. Vast numbers of animals are used for testing toxicity in the pharmaceutical and cosmetic industry every year. Nevertheless, in line with the 3Rs (Replacement, Reduction and Refinement of animal testing) principle promoted by EPAA, alternative methods, such as *in vitro* arrays and *in silico* models have been prevalently established and optimized. *In vitro* methods test toxicity by measuring the biochemical and cellular assays of chemical compounds, but experimental examinations are more challenging, in terms of efficiency and operating cost (Bowes et al., 2012). Thus, *in silico* predictions of potential toxicities in the early stages of drug development have become an efficient way of improving this long and expensive process. Machine learning algorithms are used in much of our daily life in everything from smartphone apps (e.g., voice recognition, online transaction) to self-driving cars. Drug discovery studies have profited from these developments and also the many improvements in cheminformatics and bioinformatics (Ekins, 2016). Chemoinformatics and bioinformatics have brought about beneficial insights particularly via analysis of high throughput data from drug discovery. The following chapter sections introduce the current roles of *in silico* methods in toxicity testing and propose a framework for a combined biological-chemical approach that can improve the efficiency and reduce cost for toxicity prediction.

1.1 Cheminformatics-based predictions in toxicology

Cheminformatics-based predictions broadly rely on structure-based methods. The principle is to link chemical structures (represented by molecular descriptors) to their activities or properties (e.g., toxicity endpoints). The methods broadly can be classified into QSAR (quantitative structure-activity relationship) and expert system approaches (Valerio, 2009). More precisely, QSAR modeling establishes quantitative relationships between chemical structures and activities via statistical functions (activity = *f*(structural properties), where activity might be the value of IC_{50}) (Valerio, 2009). Structural properties are represented by chemical descriptors, ranging from physicochemical properties (e.g., molecular weight, hydrophobicity, and charges), chemical fragments (e.g., structural alerts), molecular fingerprints (e.g., MACCS keys) to quantum mathematics (e.g., orbital energies) (Todeschini and Consonni, 2000). To set up this mathematical function, a set of chemical descriptors are calculated for the group of chemical compounds with known activities (e.g., IC_{50}). Descriptors are then weighted according to their correlation with activity values, thus defining the most accurate function to predict activity based on chemical structure.

With the evolution of chemical descriptors, a wide range of algorithms have been applied to QSAR, including Multiple Linear Regression (MLR), k-Nearest Neighbors (KNN), Principal Component Analysis (PCA), and Support Vector Machines (SVM) (Nantasenamat et al., 2010). Recently, within the framework of Big Data methods, a novel algorithm - Deep Neural Network (DNN) - has been developed and was shown to outperform all the other methods participated in the Tox21 Data Challenge (Unterthiner et al., 2016).

In contrast to the above algorithms, expert systems attempt to capture human expert knowledge related to structure features and their relationship to toxicity and mechanism. Methods define rules that represent more interpretable means to label compounds and relating them to potential toxicities. These structure features are known as structural alerts (SAs) or toxicophores. For example, some substructures are used as mutagenicity alerts in the expert systems DEREK (Ridings et al., 1996) and Toxtree 2.1.0 (Benigni and Bossa, 2008). Overall, chemical structure-based models can be defined quantitatively (e.g., QSAR defines mathematical function) and qualitatively (e.g., expert systems define toxicity by the presence of structural alerts), which enables chemists or pharmacists to exploit QSAR models with higher accuracy to filter drug candidates and regulation agencies to check new drugs for interpretable mechanisms of toxicity.

Most QSAR predictors don not fully capture complicated biological processes. This is sometimes due to defects in data selection, modeling approaches, and validation. There have been many attempts including training data set curation, descriptor selection, appropriate modeling algorithms development, and validation strengthen have been made to improve the QSAR modeling.

In the reality, the complex relationship between chemical structures and toxicities is likely never to be fully understood without realistic models of organismal biology. Simple models do not capture the complexity of ADME (absorption, distribution, metabolism, and excretion) process in pharmacokinetics (Hansch et al., 2004) and are also woefully inadequate to capture the even greater complexity and diversity of toxic responses. QSAR is nevertheless useful for predicting specific and direct chemical-activated toxicities (e.g., mutagenicity that could be predicted by the interactions of chemicals and DNA) even if it is less successful at predicting toxicities arising from chemical-perturbed biological pathways involving proteins targeted by chemicals (e.g., carcinogenicity) (Ashby and Tennant, 1991; Benigni and Bossa, 2008). Human rule-based expert systems also have their limitations, because the rules derived from knowledge of human experts are incomplete and biased, and updating the system is difficult, requiring extensive literature to be carefully reviewed. These limitations make accurate and efficient toxicity predictors challenging.

1.2 Bioinformatics-based predictions in toxicology

The increasing availability (and decreasing cost) of high-throughput chemical screens (HTS) and high-throughput ‘omics’ tools (transcriptomics, proteomics, etc.) has been paralleled by the development of associated computational bioinformatics tools and expertise to help unravel mechanisms of diseases and drug toxicity. Adding novel data streams to models has been proposed to help alleviate the shortcomings mentioned above.

Omics studies provide large datasets for each toxic endpoint, and thus give the system view of toxicity, involving all the molecular players (e.g., DNA, RNA, proteins, and metabolites) instead of considering each of them separately. When a biological system is perturbed by an exogenous agent (e.g., a drug), the expression of some genes may be dysregulated (transcriptomics), leading to the dysfunctions of proteins (proteomics) and metabolites (metabolomics) (**Figure 1.1**). Toxicogenomics was developed to harness the collective technologies of genomics, proteomics and metabolomics, to study the adverse effects of toxic substances on biological system and ultimately to improve efficiency and safety through a better understanding of mechanism (Hamadeh et al., 2002).

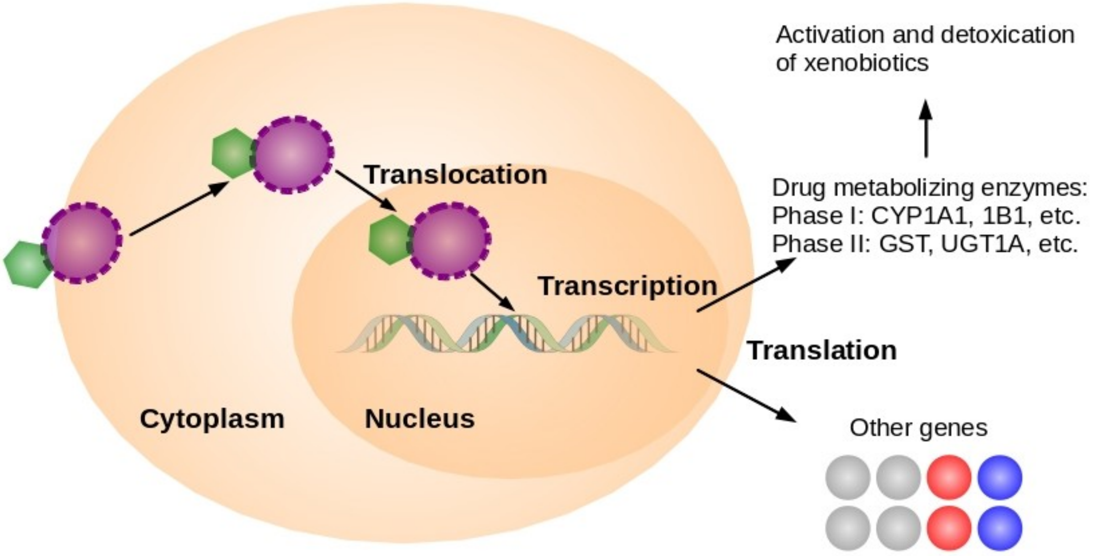

Activation and detoxication of xenobiotics
Translocation
Drug metabolizing enzymes:
Phase I: CYP1A1, 1B1, etc.
Phase II: GST, UGT1A, etc.
Transcription
Translation
Cytoplasm
Nucleus
Other genes

Figure 1.1 Illustration of gene expression changes elicited by a drug and its metabolic pathway. The drug (green hexagon) binds to a receptor (magenta ellipse) on the cell surface and a series of signaling events are triggered, resulting in the expression of drug metabolizing genes, such as phase I modification enzymes (e.g., CYP1A1 and 1B1) and phase II conjugation enzymes (e.g., GST and UGT1A), leading to activation, detoxification and excretion. Meanwhile, the toxic features of the drug cause gene expression changes (e.g., red ellipses are up-regulated and blue are down-regulated).

Transcriptomics measures the genome-wide mRNA expression (Barh et al., 2013) to provide gene expression changes involved in pathways or networks that might lead to toxic effects. DNA microarrays are one of the most common techniques. These measure tens of thousands of gene expression values by assessing the degree to which a sample of RNA or DNA hybridizes to a chip containing DNA probes (usually from a whole genome). The utility of gene expression profiles of toxicant responses is based on the presumption that compounds with similar toxicity mechanisms will give similar transcriptional changes (Waring et al., 2001). There is now a vast literature of microarrays applied to toxicology. For example, Huang et al., exploited cDNA microarrays to examine the expression alterations of genes involved in pathways of cell apoptosis, calcium homeostasis, and proposed a putative cisplatin-induced nephrotoxicity mechanism (Huang et al., 2001). Microarray studies are typically limited by the complexity of understanding gene expression changes and how they related to toxicity, and the oft-cited problem that alterations at the transcriptomic level do no always reflect protein expression (Barh et al., 2013).

Proteomics is the systematic, normally high-throughput study of proteins in a biological context (e.g. cell, tissue, organ, disease, etc.). One typically measures expression, interactions with other molecules or aspects of function (e.g. post-translational modifications). Proteomics experiments are sometimes used to complement toxicogenomics, to help assess biological response by capturing altered proteins associated with the toxicant exposure. Key technologies include 2D gel electrophoresis combined with mass spectrometry analysis for protein separation (Shruthi et al., 2016), multi-dimensional liquid chromatography/mass spectrometry for protein characterization (Shruthi et al., 2016), and fluorescence resonance energy transfer for protein-protein interactions detection (Barh et al., 2013). For example, Linge and coworkers used 2D gel electrophoresis to define fourteen altered proteins as novel biomarkers in melanoma patients (Linge et al., 2012). While several useful protein biomarkers for particular toxic responses have been identified (e.g. kidney injury molecule-1, Kim1, for nephrotoxicity) (Ichimura et al., 2008), there are still many pathologies lacking any clear biomarker. Biomarkers, once identified, can both aid the prediction of toxicities for new compounds and also help to understand the underlying mechanisms that drive particular pathophysiologies.

Metabonomics is the study of the metabolic response of a living system to external stimuli or genetic perturbation. Within toxicology, the hypothesis is that toxicant-induced pathologies lead to modifications of metabolic pathways via changes in the concentrations of metabolites. Such changes can be determined by the changes of flux distribution compared with that of the steady state (Carbonell et al., 2017). Common techniques include gas chromatography mass spectrometry (GC-MS), liquid chromatography mass spectrometry (LC-MS), and nuclear magnetic resonance (NMR) spectroscopy, all of which measure these concentrations (Wishart, 2016). One well-known application of metabonomics in toxicity prediction was performed within the Consortium for Metabonomic Toxicity (COMET) where metabonomics was used to build an expert system to predict the drug-like-induced liver and kidney toxicities in rodents (Lindon et al., 2005). More recently, even smartphone streamlined metabonomics apps for chemical toxicity assessment have also been presented (Kwon et al., 2014).

Overall, computational modeling, in the context of toxicology, has to some extent disregarded detailed information about chemical compounds, while cheminformatics-based approaches typically neglect biological mechanism. In the era of Big Data, masses of toxicity data, consisting of both chemical and biological attributes, argues for an integration of these two approaches. Clearly, toxicities of drugs and the associated mechanisms should be studied from the viewpoint of the whole system (i.e., systems toxicology).

1.3 Integrated chemical and biological modeling-based predictions in toxicology

From the viewpoint of systems biology, drug action can be seen as the perturbation of the biological systems, which involves drugs interacting with both desired targets and undesired proteins (i.e. off-target interactions), as well as associated protein-protein interactions, metabolic pathways and signaling pathways, all of which produce intended and unintended effects (Keiser et al., 2009). It is known that most drugs bind to multiple target proteins with a wide range of affinities, which together is responsible for both desired drug actions and toxicity (Walker et al., 1999). For example, a cardiovascular agent, propranolol, which targets β-adrenoceptors (i.e. a β-blocker), was observed to interact with receptors in secondary organ systems, leading to bronchoconstriction in asthma (Ahmed and Branley, 2009), characterized as an exaggerated on-target effect. Another example is the anti-obesity medication, fenfluramine, which induced valvular heart disease via activation of 5-Hydroxytryptamine receptor 2B (5-HT2B), what led to its withdrawal from the market (off-target effects) (Roth, 2007).

Overall, drug toxicity can be considered as how drugs via molecular interactions trigger adverse outcomes across different levels of human organization (cellular, tissue, organ, system) over time and as a function of dose (**Figure 1.2**). Unfortunately, only a few relationships between drug targets and toxicity endpoints have been well established, with most mechanisms remaining unknown (Vedani et al., 2012). Recent advances in big data acquisition, complemented by new experimental tools make systems toxicology, via the simultaneous combination of biological and chemical information more feasible.

Carbonell et al. have developed a systems-wide protocol for modeling organ-specific toxicity. Models integrate gene expression data, metabolic pathway signatures and hepatotoxicity observations in primary human hepatocytes treated by 77 chemical compounds. Cheminformatics models, based solely on chemical structure features, found some relationships between chemical features and toxicities. The integration with gene-expression data also provided general mechanistic hypotheses for hepatotoxicity (Carbonell et al., 2017).

Perualila-Tan and colleagues developed a joint model by integrating transcriptomic, phenotypic and chemical structure data. The model aimed to predict the impact of each chemical structure feature on gene expression and toxicity pIC50 (–log IC50) values. Two case studies were used to test the applicability of the model in drug discovery. The EGFR (epidermal growth factor receptor) dataset contained 35 compounds and expression data for 3595 genes, and the ROS1 (reactive oxygen species) dataset contained 89 ROS inhibitors with 312 unique substructures. After pooling individual predictions from the compound-gene expression and compound-substructure models into a final predicted value, the joint model gave an integrated view of how chemical substructures were associated with toxicities. For example, the model revealed that the higher expression of the genes *TXNRD1* and *FNIP1* (ROS1 inhibition and cancer related genes) were linked to the presence of the particular substructure FF-2086493472 (Perualila-Tan et al., 2016).

A relative new conceptual construct for toxicity assessment hosted by the Organisation for Economic Co-operation and Development (OECD) are adverse outcome pathways (AOPs), which capture interactions between chemical initiating events and adverse outcomes through curated biological (Ankley et al., 2010). This concept has been used, for example, to aid environmental toxicity prediction through the use of well-understood biological pathways and genomic data, which has highlighted mechanisms downstream of the initial chemical-induced key events. Algorithms analyzing data in the AOP framework are still somewhat limited. Recently, Antczak et al. built a multistep modeling method, which combines traditional QSAR modeling, gene expression and toxicity information from *Daphnia magna*. They described a strategy for defining putative AOPs via the integration of chemical structures with mechanistic insights from genomic data, and observed adverse outcomes (Antczak et al., 2015).

Stem cells have also recently played important roles in toxicity prediction. For example, Yamane and co-workers first constructed a predictive SVM model using qRT-PCR data from human embryonic stem cells (hESC) treated by twenty compounds. After adding edge weights for a gene-gene interaction network to the model, prediction accuracy for neurotoxicity, genotoxic and non-genotoxic carcinogens reached to at least 97.5%. Moreover, this study used only undifferentiated hESCs, representing the potential to predict late-onset toxicity, and the toxicities observed during embryonic development (Yamane et al., 2016).

An et al. (An et al., 2016) proposed a predictive model for assessing adverse drug reactions within liver and kidney. They trained 210 chemical features extracted from 108 drugs in Japanese Toxicogenomics Project database (Uehara et al., 2010) by four models (ANN, kNN, LDA, and SVM) to predict toxicity. Compared to other models predicting toxicity within single organs, their multiple organ models (i.e. considering organs other than the target) provided more comprehensive knowledge of drug-induced adverse effect mechanisms in humans (An et al., 2016).

1.4 Predictive toxicology at human population level

Accessing individual differences in drug toxicity is a new challenge for toxicology and personalized medicine. Accordingly, pharmacogenomics has emerged to help uncover how human genetic variation affects drug efficacy and toxicity and provide more precise prescription to patients. Genetic variations, including common genetic variants, mostly are single-nucleotide polymorphisms or SNPs, and rare variants in genes that participate in drug pharmacodynamics or pharmacokinetics play a role in individual variability (Roses, 2000). SNPs in genes encoding drug targets, drug transporters, drug-metabolizing enzymes could impact patients' sensitivity to a drug (McLeod and Evans, 2001). So far, many efforts have been put to identify associations of genetic variations and drug-induced phenotypes. Genome-wide association studies (GWAS) is one of the most popular methods for this purpose. For example, GWAS has successfully identified variants located on VKORC1 (vitamin K epoxide reductase complex subunit 1), CYP2C9, and CYP4F2, the targets of the anti-coagulant warfarin. Some variations in VKORC1 are involved with resistance to warfarin, and others affect warfarin dosage within the normal range prescribed. Variations in CYP2C9 and CYP4F2 are associated with the speed of metabolizing warfarin. In poor metabolizers (slow-processing), warfarin blood concentration can rise, leading to toxicity (Takeuchi et al., 2009). With the help of GWAS, a great number of SNPs related phenotypes have been identified. Recently, technological advances in sequences mean that exome or even whole genome sequences are more commonplace (The 1000 Genomes Project Consortium, 2015; Lek et al., 2016). The 1000 Genomes (The 1000 Genomes Project Consortium, 2015) and ExAC projects (Lek et al., 2016), which sequenced or collected genomes of thousands of people, provide information to help fine-map pharmacodynamics and pharmacokinetics-associated loci.

Personalized or precision medicine aims to use the sequences of an individual's genes to make more rational treatment and dose suggestions for better efficacy and reduced toxicity. For example, Roche has developed the AmpliChip to detect the cytochrome P450 gene family variants that might affect drug efficacy or detoxification. 35% of the general population has found to carry abnormal CYP2D6 alleles, making them either poor or ultra metabolizers. AmpliChip can detect up to 33 CYP2D6 alleles to aid treatment decisions (Rebsamen et al., 2009). Computational phenotyping facilitates a fast development of quantitative models for individual's drug toxicity risk. For instance, integrating preclinical biological experiments, clinical, human imaging and drug information, allowed the prediction of clinical responses for certain drugs (Geerts et al., 2015). Overall, personalized medicine is clearly an emerging theme in healthcare, and a wealth of novel discoveries from this field will ultimately improve risk assessment in medicine (Ingelman-Sundberg, 2015).

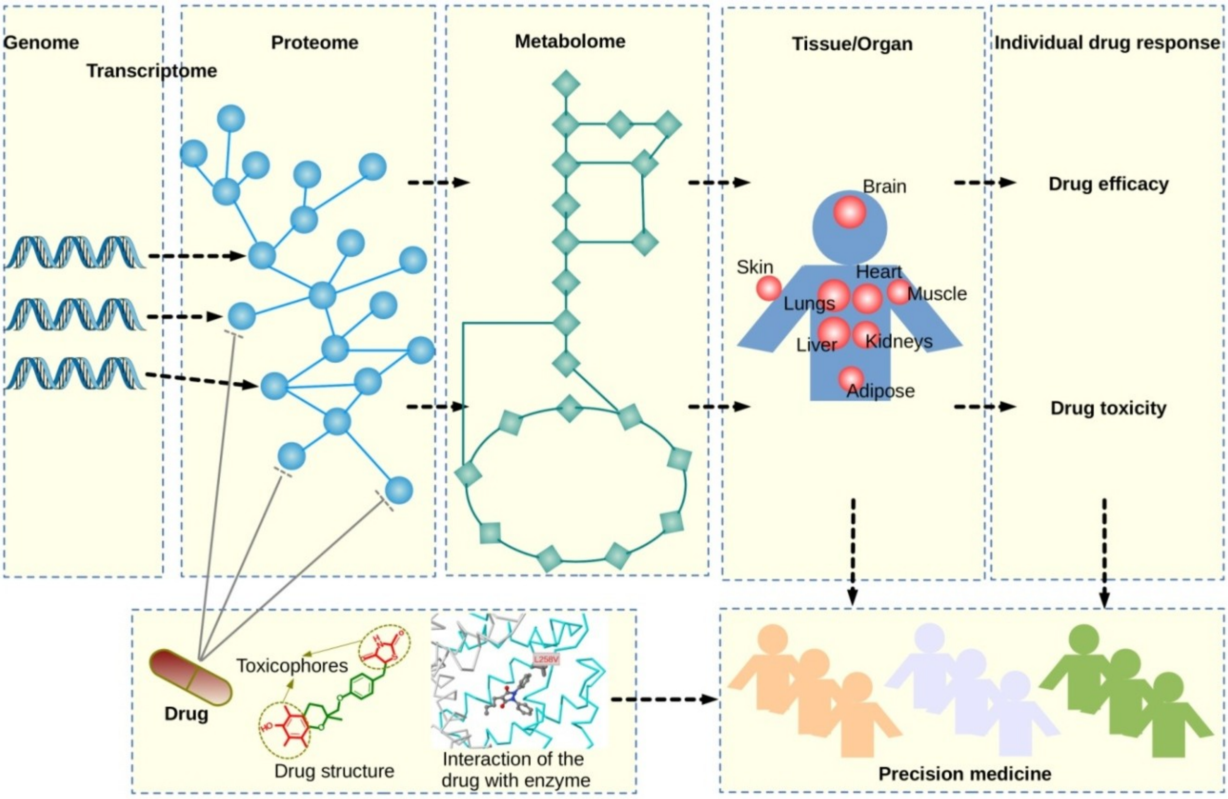


Figure 1.2 Scheme of drug toxicity prediction in complex biological systems. Drug chemical structures, targets, SNPs (target or other genes), the associated biological networks, metabolic pathways, tissues/organs influenced by the drug are all keys to understanding and predicting biological responses.

1.5 Dissertation outline

Technologies such as gene sequencing, transcriptomic profiling, proteomics, metabolomics, and pre-clinical or clinical data related to single cells, organs, or patient populations provide the Big Data relevant to systems pharmacology and toxicology, which in principle should allow better predictions of drug action and adverse responses. Only when all these data are integrated can researchers and clinicians develop models or concepts to better predict outcomes.

This dissertation presents systems approaches to addressing some of the above problems (**Figure 1.2**). Chapter II presents a chemical fragment/substructure-based approach to toxicity prediction exploiting machine learning (support vector machines (SVM) and sparse canonical correlation analysis (SCCA)) methods. Such fragment-based approaches are well-established for the development of novel lead compounds (Schuffenhauer et al., 2005). However, they can also be employed to identify fragments indicative of toxicities (Siegel and Vieth, 2007). We propose a predictive toxicity protocol based on chemical structures. We focus initially on the systematic derivation of chemical fragments associated with particular toxicities using an overall literature basis and avoiding any bias or limitations introduced by expert curation. Some of the enriched fragments for specific toxicities are potential structural alerts for future predictions on new drugs. We also provide mechanistic interpretation for some of the fragments by integrating chemical-protein interactions, protein-protein interactions and gene expression data.

Chapter III generally explores the mechanisms of predictive toxicity. Predictive systems are generally more accepted if they provide descriptions of mechanism (e.g., drug metabolizing pathways) that underlie the biological properties of the molecule. The response of a biological system to a toxicant that afterwards causes pathology in certain organs can then be examined as changes in the expression of genes, proteins synthesis, and metabolism. Of these changes, the expression of genes is the most sensitive and readily accessible experimentally. Therefore, toxicogenomics, which measures gene expression changes caused by a toxin in a specific cell, tissue or organ, has become one of the most powerful strategies. We employ the Toxicity Evaluation System developed by the TGP in Japan (Uehara et al., 2010), which provides comprehensive toxicogenomics data for hundreds of compounds to uncover relationships between changes in gene expression and toxicity data.

Resources like this can help to identify candidate gene/protein biomarkers of toxicity. Such datasets also permit us to test hypotheses such as whether or not *in vitro* bioassays can be utilized to predict *in vivo* responses.

Chapter IV attempts to study the underlying mechanisms of drug response differences among individuals and populations. Besides chemical and bioassay results, the recent availability of data for healthy individuals provides new possibilities to study drug effects in human populations. The Mechismo tool (developed in our group) provides potential mechanisms for how proteins interact with other molecules (i.e., proteins, chemical compounds, nucleotides) and how any changes/variations might affect these interactions and consequently an entire biological system (Betts et al., 2015). In this chapter, we use Mechismo to study data from the 1000 Genomes Project to investigate which genetic variations might impact drug response differences among individuals and subpopulations.

The concluding chapter summarizes the findings from the entire thesis and discusses their contributions towards predictive toxicology and personalized medicine.

CHAPTER II: Chemical fragments as foundations for understanding the toxic effects of chemicals on biological systems

2.1 Abstract

The building blocks (fragments) of chemicals can sometimes be associated with particular biological effects of the chemical as a whole (e.g., metabolism, disease, pharmacological actions, toxicity). In this chapter, we present a systematic strategy to predict and understand the mechanistic interpretation of chemical toxicities based on chemical fragments.

To identify toxicity structural alerts (SAs), we derived 93 902 chemical compounds with structures and MeSH (Medical Subject Headings) terms from PubChem (Bolton et al., 2008), and their fragments from the ZINC database (Irwin and Shoichet, 2005). We classified compounds into different toxic (or non-toxic) classes according to information extracted from the literature, and performed statistical analyses to identify chemical fragments enriched particular classes (as potential structural alerts).

Hundreds of fragments strongly related to specific toxicities were identified, which might be regarded as structural alerts for using in toxicity prediction. For validation, we performed toxicity predictions with support vector machines (SVM) and sparse canonical correlation analysis (SCCA) for 263 withdrawn drugs. Some predictions were supported by mechanistic explanations obtained by integrating fragment-chemical, chemical-protein, protein-protein interactions and gene expression data. Our findings might ultimately help understand toxicities of unknown chemicals and their potential mechanisms to aid drug development.

2.2 Introduction

Several *in silico* methods for predicting toxicities have been proposed previously. Most approaches relate particular chemical fragments to toxicity by statistical analysis, meaning that the fragment is significantly enriched in toxic compounds compared to others (Schnur et al., 2006). To date, a number of structural alerts (toxic fragments/ toxicophores that are associated with toxicity) (Ashby and Tennant, 1988)) have been defined for several toxic endpoints (Liu et al., 2015). For example, Ashby and Tennant identified a set of structural alerts correlated with the modification of DNA based on the data from both *in vivo* carcinogenicity and *in vitro* mutagenicity (Ashby and Tennant, 1991). Benigni and Bossa reported 31 classes of structural alerts for carcinogenicity and mutagenicity with corresponding mechanisms of action (Benigni and Bossa, 2011), which has been implemented as prediction rules into Toxtree 2.1.0 (Benigni and Bossa, 2008). Expert systems are either developed manually or statistically by extracting expert knowledge, to predict the molecular toxicity based on the presence or absence of toxic fragments in their chemical structure. The program DEREK (from Lhasa Limited; http://www.lhasalimited.org) is an example of rule-based expert system where most general endpoints (e.g. reproductive toxicity and genotoxicity) are directly indicated by toxic fragments like phthalate, or nitrobenzene, if they have been detected in the examined molecules (Ridings et al., 1996). Another example is HazardExpert from CompuDrug, where the toxicity prediction is based on the list of known toxic fragments collected from the U.S. Environmental Protection Agency (Brink and Walker, 1987). Similarly, CASE, probably the first automated expert system, automatically recognizes structural alerts in new chemicals (Klopman, 1984). Other programs such as PASS (Poroikov et al., 2000), Cat-SAR (Cunningham et al., 2008), and LAZAR (Helma et al., 2004) also identify structural alerts indicative of toxicities. Rule-based systems have some limitations (Lepailleur et al., 2013; Liu et al., 2015), specifically: i) the number of structural alerts is limited for general toxic endpoints (e.g., hepatotoxicity) ii) they require extensive man-power to harvest results from the literature to remain up to date, and iii) the opinions of the experts can lead to biases and potentially inaccurate results i.e., non-toxic compounds containing a structural alert might be predicted to be toxic.

The development of data mining tools enables one to detect large-scale novel structural alerts beyond the limits of human perception. Commonly used machine learning algorithms for this purpose include multiple linear regression (MLR), partial least squares (PLS), support vector machines (SVM) and k-nearest neighbors (k-NN). For example, in a study from Zhang et al., liver toxicities of 1317 compounds were predicted using machine learning methods (e.g., SVM, k-NN, Random Forest, RF) based on a substructure pattern recognition method (Zhang et al., 2016). The best model (SVM) showed much higher predictive accuracies for the training set, test set and an external validation set compared to previous QSAR methods. Moreover, six structural alerts related to the mechanism of drug-induced liver injury (DILI) were identified, which could be utilized for irradiating chemicals that might induce DILI. Mazzatorta et al. reported the correlation of chemical substructures from over 400 compounds and the chronic toxicity by a predictive model built on MLR and PLS methods. The model predicted that the chronic toxicity is caused by compounds containing particular chemical moieties (Mazzatorta et al., 2008). Pauwels et al. proposed sparse canonical correlation analysis (SCCA) model to predict potential side-effects of drugs in large molecular database based on their chemical fragments. The unique feature of this model is the capacity to extract the associations of side-effects of drugs and the emergence of structural alerts (Pauwels et al., 2011). The software SARpy automatically generated and selected chemical fragment structural alerts by analyzing the correlation between the frequency of each chemical fragment and the experimental activity of the chemicals (Ferrari et al., 2013).

Predictive systems are more acceptable if they provide a description mechanism, i.e., the interactions between the molecule and target proteins or systems that lead to its ultimate biological effects. The response of a biological system to a toxicant that afterwards causes pathology in certain organs can be examined as changes in the expression of genes, proteins synthesis or metabolism. Of these, gene expression is the easiest to measure and interpret. For example, Low et al. developed hybrid models combining chemical and toxicogenomics descriptors for 127 drugs from the Japanese Toxicogenomics Project (Uehara et al., 2010) using classification methods (k-NN, SVM, RF, and distance weighted discrimination, DWD). Besides identifying and verifying chemical structural alerts for hepatotoxicity by HiT QSAR (Kuz'min et al., 2008) and XCHEM (Sedykh and Klopman, 2006), the transcripts predicted related to DILI mechanisms were also predicted (Low et al., 2011). Hewitt et al. clustered 951 diverse compounds into a few structurally clusters by molecular structure similarity and identified sixteen structural alerts associated with human hepatotoxicity, which were supported by mechanistic insights by gene expression (Hewitt et al., 2013). Antczak et al. proposed a system level strategy to define putative toxicity pathways by integrating chemical structure information with pathways defined by gene expression profiles induced by 26 chemicals, and their observed phenotypic effects (Antczak et al., 2015).

An increasing number of databases are available to aid drug discovery by providing information on chemical fragments, chemical-protein interactions, protein-protein interactions, toxic endpoints and gene expression. For example, the ZINC database stores 162 261 (2013-11-05) “fragment-like” molecules that have molecular weight less than 250 g/mol, LogP values between -2 and 3, fewer than three hydrogen-bond donors, fewer than six hydrogen-bond acceptors and fewer than three rotatable bonds (Irwin and Shoichet, 2005). FragmentStore uses modified Lipinski rule-of-five criteria (Lipinski et al., 2001) to validate fragments for assembling compound libraries. The library contains more than 35 000 fragments derived from more than 13 000 metabolites, 2 200 toxic compounds and 16 000 drugs (Ahmed et al., 2011).

There are also a wide range of toxicity databases capturing different measures of toxicity. The ToxCast program (Kleinstreuer et al., 2014) and Tox21 (Mahadevan et al., 2011) developed by the United States Environmental Protection Agency (US EPA) provide overall toxic effects for about 10 000 environmental chemicals and drugs. Non confidential regulatory submissions such as the European Chemicals Agency (ECHA) (http://echa.europa.eu/) and INCHEM (http://www.inchem.org/) also provide relevant data of toxic studies on compounds.

In order to link the chemicals to proteins, interactions from a variety of databases of chemical-protein interactions are available, for example, experimental evidence interactions from protein data bank (PDB) (Gutmanas et al., 2014) and PDSP K_i databases (Roth et al., 2000), interactions from pathway databases KEGG (Kanehisa et al., 2012) and Reactome (Fabregat et al., 2016), and interactions from integrated database such as STITCH (Kuhn et al., 2014) which merge other databases. Resources that archive published protein-protein interactions are also available, such as Network Database (BIND) (Bader et al., 2003), Biological General Repository for Interaction Databases (BioGRID) (Chatr-aryamontri et al., 2015), the Database of Interacting Proteins (DIP) (Xenarios et al., 2002), IntAct, Molecular Interaction database (MINT) (Orchard et al., 2014). All of these are merged in the Proteomics Standard Initiative Common Query Interface (PSICQUIC) (Aranda et al., 2011).

Toxicogenomics datasets of gene expression (e.g. CMAP (build 02) (http://www.broad.mit.edu/cmap/) (Lamb et al., 2006), as discussed in the last chapter, also provide details on expression changes induced by toxicants.

In this chapter, we propose a combined method to predict toxic endpoints. The method exploits a combination of statistical analyses with machine learning algorithms (SVM and SCCA) to investigate the correlation between compound fragment with toxic endpoints derived from databases and the literature. We achieved good performance and high accuracy for predicting toxic endpoints of chemicals with proposed structural alerts. We also supplement prediction with putative mechanisms of toxicity, if possible, data on chemical-protein interactions (both on and off-targets), protein-protein interactions and gene expression datasets used to make the predictions.

2.3 Methods

2.3.1 Data preparation

93 902 chemical compounds with MeSH terms and their structure files were downloaded from PubChem (2013) (Bolton et al., 2008), and chemicals with fewer than four non-hydrogen atoms and more than 20 atoms were removed.

The external validation set containing 258 known toxic compounds with annotated gene expression changes in toxicogenomics experiments was retrieved from ToxWiz (http://www.toxwiz.com/).

263 withdrawn drugs as the test set to evaluate the interest of the models for predicting toxicities for uncharacterized drugs were collected from WITHDRAWN database (Siramshetty et al., 2016).

The clean fragment-like subset of 162 261 molecules dated on November 05, 2013 was downloaded from ZINC (version 12) (Irwin and Shoichet, 2005) as the substructure dataset.

All compounds were represented by SMILES strings, a widely used notation of encoding chemicals as ASCII strings (Weininger, 1988). Additional refinements including adding hydrogen to fulfill the valences of non-hydrogen atoms and neutralizing the charges were processed for all the chemical files with Open Babel 2.3.2 (O'Boyle et al., 2011). Conversions of files from .mol, .sdf chemical format to SMILES and the substructure searches were also implemented using Open Babel 2.3.2.

2.3.2 Extracting toxicity information

A total of 20 toxicity categories were defined based on the pathology pathways in ToxWiz (http://www.toxwiz.com/). These categories include 1504 MeSH terms of toxicity endpoints from 20 systems: male reproductive system, development, digestive system (excluding hepatic), lymphatic system, endocrine system, sensory apparati, integumentary system (e.g. skin, hair), activity system (e.g. muscle, bone), connective and other tissues & cells, pan-systemic pathologies, hepatic system, urinary system, circulatory system, respiratory system, immune system, nervous system, reproductive system (general), other toxicity related clusters, female reproductive system. The most prevalent toxicities (in terms of numbers of compounds) are neurotoxicity, hepatotoxicity, cardiotoxicity and genotoxicity.

Compound-toxicity relationships were generated by automatically identifying PubChem-MeSH-PubMed links and crossing compounds with references containing the textual terms (e.g. hepatotoxicity, carcinogenicity, mutagenicity, developmental toxicity).

2.3.3 Classifying fragments and dataset diversity

An important aspect of the fragment-based toxicity prediction is to assess diversity within a set of compounds having a particular property. This is important to avoid highly similar compounds (i.e. those that are likely derived from one another) dominating the fragments groups. A Tanimoto similarity calculation, for fragment dataset, was performed by a method based on MolPrint 2D (Bender et al., 2004) to determine whether a fragment is structurally novel. The Tanimoto coefficient (TC) is defined as follows:

$$T(a,b) = \frac{Nc}{Na+ Nb-Nc} \tag{2.1}$$

where N_a and N_b are the number of bits set (denoting presence or absence of a particular fragment) for binary fingerprints of molecules A and B, respectively, and N_c is the set bits that A and B have in common. To estimate whether a dataset is structurally novel, an all-by-all similarity matrix was calculated for these compounds. We defined compounds to be similar if the Tanimoto coefficient was 0.85 or above (Martin et al., 2002). We left single representatives for each group, and removed groups having fewer than four compounds.

2.3.4 Toxic fragments analysis

In order to identify whether a specific fragment is more frequent in a toxic than a nontoxic class, the ratio of the observed/expected value was calculated. If a fragment was more frequently presented in a particular toxic class than the expected number, this fragment was determined to be toxic. The expected number of the presence of a fragment in a particular toxic class is defined as following:

$$exp = \frac{N_{fragment\_total} * N_{toxicity\_total}}{N_{total}} \tag{2.2}$$

Where $N_{fragment\_total}$ is the number of compounds containing the fragment; $N_{toxicity\_total}$ is the number of compounds in each toxic class; and N_{total} is the total number of fragment-chemical-toxicity pairs. The test of statistical significance for the observed/expected can be obtained in R *fisher.test* (Agresti, 2002). In this study we consider $p < 0.05$ is statistically significant.

2.3.5 Prediction methods

2.3.5.1 Support vector machines (SVM)

Support vector machines (SVM) are a widely used technique for classification in bioinformatics and chemoinformatics (Guyon et al., 2002; Pauwels et al., 2011). The kernlab package in R (version 0.9-24) (Karatzoglou et al., 2004) was used in our study. We tested linear kernel and multiple nonlinear kernels; the Gaussian RBF kernel function was found to have the best performance. For a good generalization capacity of the classifier, two kernel parameters should be defined well, i.e. C and σ, where C is a cost factor that balances the trade-off between margin and training error and σ is a kernel parameter. 5-Fold cross-validation was applied to train nonlinear SVM(s) with various C values in the range of 2^{-10} to 2^{15} and an automatically determined σ, and the combination of C and σ that generated the highest cross-validation accuracy was selected for training. The trained models were applied to the test set to predict whether a given compound is linked to toxicity endpoint or not. We constructed classifiers for each of the toxicity endpoints in our study.

2.3.5.2 Sparse canonical correlation analysis (SCCA)

Canonical Correlation Analysis (CCA) makes it possible to find linear combinations of two sets (p fragments and q toxic effects) generated from n samples (compounds) with the highest correlations. CCA enables one to find linear combinations of all the variables in matrix **X** (p × n) that maximally correlate with linear combinations of all the variables in matrix **Y** (q × n). Considering two linear combinations as canonical variates $\alpha = \mathbf{X}u$ and $\beta = \mathbf{Y}v$, with the weight vectors u' = (u_1, ..., u_p) and v' = (v_1, ..., v_q). The optimal weight vectors are obtained by maximizing the canonical correlation coefficient between the variate pairs:

$$\rho = corr\ (u, v) = \frac{v'Y'Xu}{\sqrt{v'Y'Yv}\sqrt{u'X'Xu}} \tag{2.3}$$

In regular CCA, the weight vectors u and v are not unique if p or q is large. In our study, we used Sparse CCA (SCCA) based on the method proposed by Pauwels et al. (Pauwels et al., 2011) for high-dimensional settings, where assuming that the columns of **X** and **Y** are uncorrelated, therefore, SCCA considers solving another form of CCA with the following optimization problem:

$$\underset{\alpha, \beta}{argmax}\ \alpha' \textstyle\sum XY\beta \text{ subject to } \|\alpha\|_2^2 = 1, \|\beta\|_2^2 = 1, \|\alpha\|_1 \le c1\sqrt{p}, \|\beta\|_1 \le c2\sqrt{q} \tag{2.4}$$

where $\sum$XY is the covariance matrix of **X** and **Y**, and $\|.\|_1$ is L_1 norm (the sum of all absolute values in the vector), and c_1 and c_2 are parameters to penalize the sparsity in α and β. For simplicity, in this study, $c_1 = c_2$ and $0 < c \le 1$.

The penalized multivariate analysis was performed with the "PMA" library in R (Witten et al., 2009).

2.3.6 Performance evaluation

We ran 5-fold cross-validation to evaluate the performance of the SVM and SCCA models. All models were assessed by counting the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and calculating overall predictive accuracy (accuracy = (TP+TN)/(TP+TN+FP+FN)), sensitivity (TP/(TP+FN)), and specificity (TN/(TN+FP)) (Baldi et al., 2000). Additionally, the receiver operating characteristic (ROC) curve, which is a graphical plot of the true positive rate (sensitivity), against false positive rate (1-specificity) (Gribskov and Robinson, 1996) was used to evaluate the performance, and the area under the curve (AUC) was also calculated with ROCR package in R (Sing et al., 2005) where AUC = 1.0 represents a perfect classifier and AUC = 0.5 means a worthless test (Lobo et al., 2008).

2.4 Results

2.4.1 Datasets

2.4.1.1 The fragment dataset

We used a dataset of 30 086 compounds with between 4 and 20 non-hydrogen atoms out of a total of 93 902 with MeSH terms in PubChem (2013). After cleaning charges and eliminating duplicates, 148 310 out of 162 261 fragments from the clean fragment-like subset of ZINC were used as the set of substructures, of which 11 011 were observed in a total of 26 644 PubChem compounds using substructure search in Open babel 2.3.2 (O'Boyle et al., 2011).

It is clear that many toxic endpoints can be caused by very different compounds (Benigni, 1991). However, it is still likely that highly similar structures can cause similar toxicities, and thus this situation could bias the results. Thus to further test the chemical diversity of the fragment data set, we computed the Tanimoto similarity index of the whole data set using MolPrint2D (Bender et al., 2004). From the distribution patterns of molecular similarity values among these 11 011 fragments (**Figure 2.1**), it is clear that most (~99%) of the values range from 0 to 0.4 (i.e. virtually no similarity in structure), suggesting that our data set is structurally diverse.

2.4.1.2 The toxicity endpoint dataset

The entire dataset contains 5 605 compounds linked to 1 477 toxicity endpoints, with a total 54 353 compound-toxicity associations and 115 639 fragment-toxicity associations.

2.4.2 Toxic fragment analysis

To understand the chemical fragments related to toxicities, frequency analysis was applied to recognize the structural alerts. Details of frequencies of each fragment occurred in different toxic classes with Obs/Exp > 1 and $p < 0.05$ were shown in **Table S2.1** of

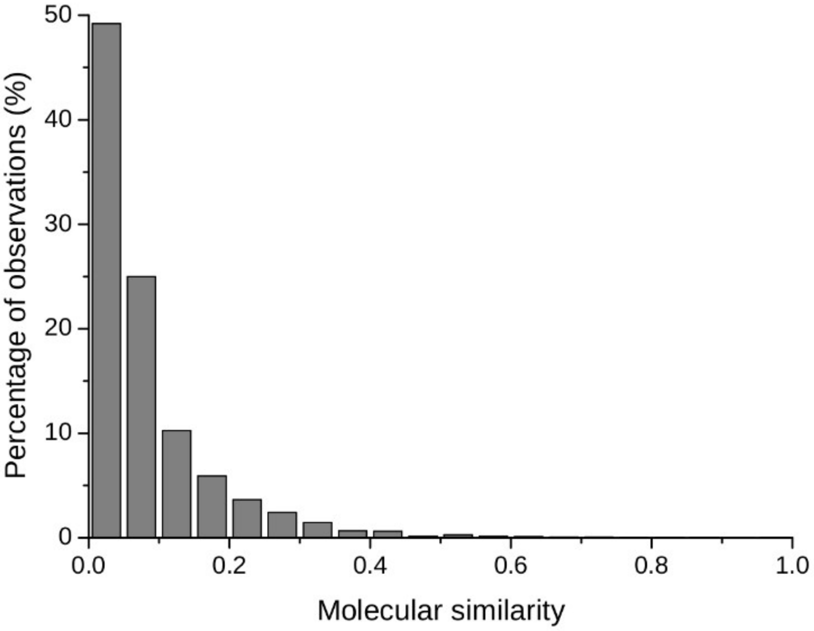
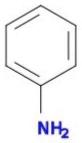
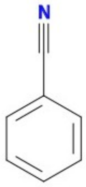
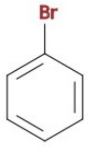
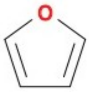
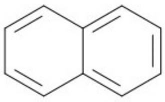
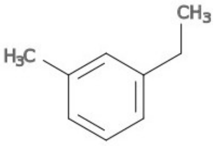
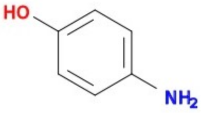
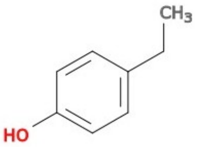
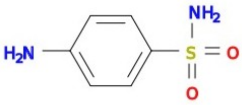
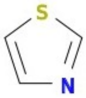
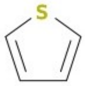


Figure 2.1 Histogram of the distribution of the molecular similarities between all fragment pairs

Supplementary Materials. These 609 fragments reflected the chemical features of 234 types of toxicity and can be considered as potential structural alerts to predict toxicity of uncharacterized compounds. **Table 2.1** shows some well-known structural alerts that were detected in our study and their statistical performance.

Table 2.1 Some examples of structural alerts

Name	Structure	Endpoints	Obs/exp	P value
Aniline		hepatotoxicity nephrotoxicity	1.40 1.37	0.0173 2.67e-05
Benzonitrile		genotoxicity	3.96	0.0009
Bromobenzene		nephrotoxcity	2.20	0.0020

Furan		nephrotoxicity	1.76	0.0105
		teratogenicity	1.78	0.0009
		genotoxicity	2.72	1.23e-09
Naphthalene		nephrotoxicity	3.12	5.07e-05
		teratogenicity	1.75	0.0039
		genotoxicity	3.26	3.61e-20
3-Ethyltoluene		cardiotoxicity	1.73	0.043
		nephrotoxicity	1.34	0.014
4-Aminophenol		hepatotoxicity	2.76	0.0001
		teratogenicity	1.57	0.0010
		nephrotoxicity	1.62	0.0009
4-Ethylphenol		cardiotoxicity	2.52	0.0009
Sulfanilamide		epidermis necrosis	24.67	2.05e-08
Thiazole		neurotoxicity	1.96	0.0015
Thiophene		hepatotoxicity	8.31	0.0004
		genotoxicity	2.10	0.0008

Several of the alerts identified correspond to compounds or moieties already associated with toxicities. For example, the aniline moiety, a common structural alert, is found in many drugs (Kalgutkar, 2015). Aniline groups can be oxidized to intermediates and transformed to phenylhydroxylamine and nitrosobenzene groups, which in turn can covalently bind macromolecules or radicals to induce toxicities (Kalgutkar, 2015).

Benzonitrile is contained in some pesticides, such as ioxynil, chloroxynil, and bromoxynil, and it is reported that these compounds induce severe toxicity in human cells (Lovecka et al., 2015).

Bromobenzene-induced nephrotoxicity is thought to occur after bromobenzene is first metabolized to 2-bromophenol and then to bromohydroquinone, which is then transported to the kidney where it induces toxicity (Schnellmann and Mandel, 1986). It is also thought to induce hepatotoxicity via GSH depletion, reduction of nucleophiles such as NAD(P)(H) and CoA, and by chemical intermediates binding to proteins (Schnellmann and Mandel, 1986).

Furan related compounds are involved in a wide range of toxicities (e.g., genotoxicity, nephrotoxicity and teratogenicity) and some mechanisms have been investigated (Moro et al., 2012). For example, furan could be activated by cytochrome P450 enzymes into reactive compounds such as α,β-unsaturated dialdehyde, cis-2-butene-1,4-dial, which can bind to nucleophiles to form cytotoxic adducts (Byrns et al., 2002). Another study showed that cis-2-butene-1,4-dial forms glutathione conjugates and protein adducts to amino and thiol groups of amino acids, leading to toxicity in the target organs (Chen et al., 1997).

Naphthalene rings are implicated in various toxicities. The mechanism of naphthalene-induced nephrotoxicity is first the metabolism by CYP450s into naphthalene epoxide, which induces serial reactions including GSH depletion, and naphthol or dihydrodiol production. The subsequent conversion to naphthalenediol and oxidation to naphthoquinone generates reactive oxygen species (ROS), leading to cell damage or death (Stepan et al., 2011).

4-Aminophenol is known to be activated by liver enzymes into a quinoid structure (Fowler et al., 1991), which might contribute to the potential nephrotoxicity of compounds containing this.

4-Ethylphenol induces toxicity by first binding macromolecules via the free electrons from oxidized substrates (e.g., phenoxy radicals, semiquinones and quinine methide) leading to additional ROS (Thompson et al., 1995). The acute toxicity of 4-ethylphenol has been assessed by a read-cross approach (Mellora et al., 2017), which correctly identified analogues of the 4-ethylphenol group as belonging to the same toxicity category. A literature review suggests the toxicity of substituted phenols is correlated with the formation of free radicals by abstracting a hydrogen radical from the phenolic hydroxyl group ultimately leading to cellular damage (Hansch and Gao, 1997).

Because the thiophene moiety is more sensitive to S-oxidation, thiophene-containing drugs tend to be metabolized to reactive S-oxides in the liver where they can act as inhibitors of CYP450 enzymes (Liu and Uetrecht, 2000).

2.4.3 Prediction via support vector machines

We then evaluated the capability of using the proposed structural alerts to predict uncharacterized compounds by machine learning methods. 3,743 compounds containing the 608 structural alerts (toxic fragments) and linking to the 234 toxic MeSH terms were found in our data set. Since the imbalance (the proportion of toxic and non-toxic cases) of the data set affects the learning process (Afzal et al., 2013), we reduced the classifiers to balance the positive (toxic) proportions increasing classifiers accuracy. This was achieved by removing the toxic terms that appeared in less than ten compounds.

Having assessed which fragment or toxicity is present in any compound, we then defined a binary matrix for each compound where 1 indicates presence and 0 indicates absence of each fragment or toxicity. This led to a data set comprising 991 compounds containing 415 proposed toxic fragments, and linked to 232 toxic terms, meaning a total of 647 binary values for each compound (Supplementary Materials).

Overall, we have then the potential to derive individual SVM classifiers for 232 different toxicity endpoints. We firstly applied 5-fold cross-validation support vector machines (SVM) to test the ability of predicting known toxicities by proposed structural alerts. The matrix was randomly divided into five roughly equal sized subsets, and then each subset was taken in turn as a test set, and the remaining four data sets as training sets to perform the prediction.

We evaluated the performance by accuracy (correctly predicted endpoints / total endpoints predicted) and ROC curves (plots of true positive versus false positive rates). Due to the imbalance of the dataset, the sensitivity (true positive rate) and specificity (true negative rate) were also used to measure the overall prediction performance. Statistical results of these values were shown in **Table S2.2** of Supplementary Materials. Parameters were chosen by using the AUC score as an objective function. The best result was gained by Gaussian RBF kernel when width parameter $\sigma = 0.1$ and regularization parameter $C = 1$.

Cross-validation, provided it is done correctly, will show reliable accuracy for the predictors. However, ultimate tests come from data external to that within the system. To do this, we chose the best model trained by SVM and tested with the external validation set from ToxWiz which contained 200 compounds relating with known toxic endpoints, e.g., hepatotoxicity, neurotoxicity, nephrotoxicity and genotoxicity to evaluate the predictive ability of external data. Since the external set was independent from the training and test set, the performance on the external set could reflect predictive capability. The results are shown in **Table S2.2** and **Table 2.2.** The best results were at the highest accuracy of 0.894, 0.890, 0.874, 0.865 for predicting hepatotoxicity, neurotoxicity, nephrotoxicity and genotoxicity, respectively, which represented credible predictions. However, as **Table S2.2** shows, most specificity values are much higher than sensitivity values, which means that the models have higher accuracy for predicting non-toxicities than toxicities. One reason might be that most toxic endpoints (e.g., splenotoxicity, scalp inflammation, cone degeneration) are rarely occurring during clinical trials. By contrast, the high-throughput screening methods identified many hepatotoxicity, nephrotoxicity, genotoxicity drug candidates at the preclinical stage. Therefore, the models are more capable to predict most common observed toxicities than rarely observed toxic terms.

Table 2.2 Performances of some toxic endpoints based on SVM method

Toxic endpoint	Data set	Accuracy	Sensitivity	Specificity	AUC
genotoxicity	Training set	0.865	0.904	0.828	0.934
	Validation set	0.740	0.600	0.751	0.734
hepatotoxicity	Training set	0.894	0.556	0.983	0.971
	Validation set	0.685	0.467	0.935	0.756
nephrotoxicity	Training set	0.874	0.747	0.954	0.948
	Validation set	0.895	0.533	0.930	0.953
neurotoxicity	Training set	0.890	0.896	0.883	0.953
	Validation set	0.755	0.792	0.750	0.822

We also performed predictions on 263 withdrawn drugs from the WITHDRAWN database (Siramshetty et al., 2016). 195 withdrawn drugs contained potential toxic fragments and a total of 117 drugs with seven types of toxicity were predicted based on these fragments by SVMs model (**Table 2.3**). Hepatotoxicity (103 cases) was the most commonly predicted toxic endpoint. Rare drugs were predicted to be neurotoxic, genotoxic, cardiotoxic, nephrotoxic, teratogenic and skin toxic. The remaining 78 drugs were not predicted to associate with any specific toxicity. This is likely because, firstly, the average accuracy of our models is limited to 0.87. Secondly, post-marketing drugs were recalled from the market due to various events, ranging from safety reasons to non-safety problems including inefficiency, manufacturing, regulatory concerns and business issues

(Siramshetty et al., 2016). Less than 10% of products were withdrawn worldwide and many withdrawals occurred unpredictably in one or two countries due to idiosyncratic drug reactions (Hussaini and Farrington, 2007).

Table 2.3 Predicted toxic endpoints associated with withdrawn drugs

Toxic endpoint	Number of withdrawn drugs	Main structural alerts
cardiotoxicity	6	1-butanol; 1-pentanol; 1-pentene; 2,2-dimethylpentane; 2-methylbutane; acetic acid; cyclohexane; tetraisopropoxytitanium; trans-2-pentene
genotoxicity	7	aniline; ethylbenzene; chlorobenzene; pyridinium; chlorochromate; thiophene
hepatotoxicity	103	4-aminophenol; acetic acid; aniline; anisole; bromobenzene; chlorobenzene; ethylbenzene; furan; P-cresol; phenetole; phenol; thiazole; thiophene; sulfanilamide
nephrotoxicity	3	acetic acid; aniline; naphthalene; propionamide; thiophene
neurotoxicity	10	1,6-hexanediol; 1-butanol; 1-pentanol; ether; ethylbenzene; tetraisopropoxytitanium
skin toxicity	1	aniline; ethylbenzene
teratogenicity	2	ether

Figure 2.2 shows some details for a selection of predictions. Many of these drugs are metabolized by CYP450s to form reactive metabolites, available of GSH depletion and binding to cellular proteins leading to toxicities and ultimate withdrawal from the market. For instance, NADPH-dependent covalent binding in human hepatocytes has been found in the metabolites of the uricosouric drug benzbromarone (McDonald and Rettie, 2007). Elsewhere, monohydroxylation catalyzed by CYP3A4 on the aniline moiety in the analgesic and antipyretic drug acetanilide apparently causes the formation of some intermediates that can be further metabolized to reactive quinone-imine species (Stepan et al., 2011). Ticrynafen can form reactive intermediate via epoxidation or S-oxidation of its thiophene fragment catalyzed by CYP450s.

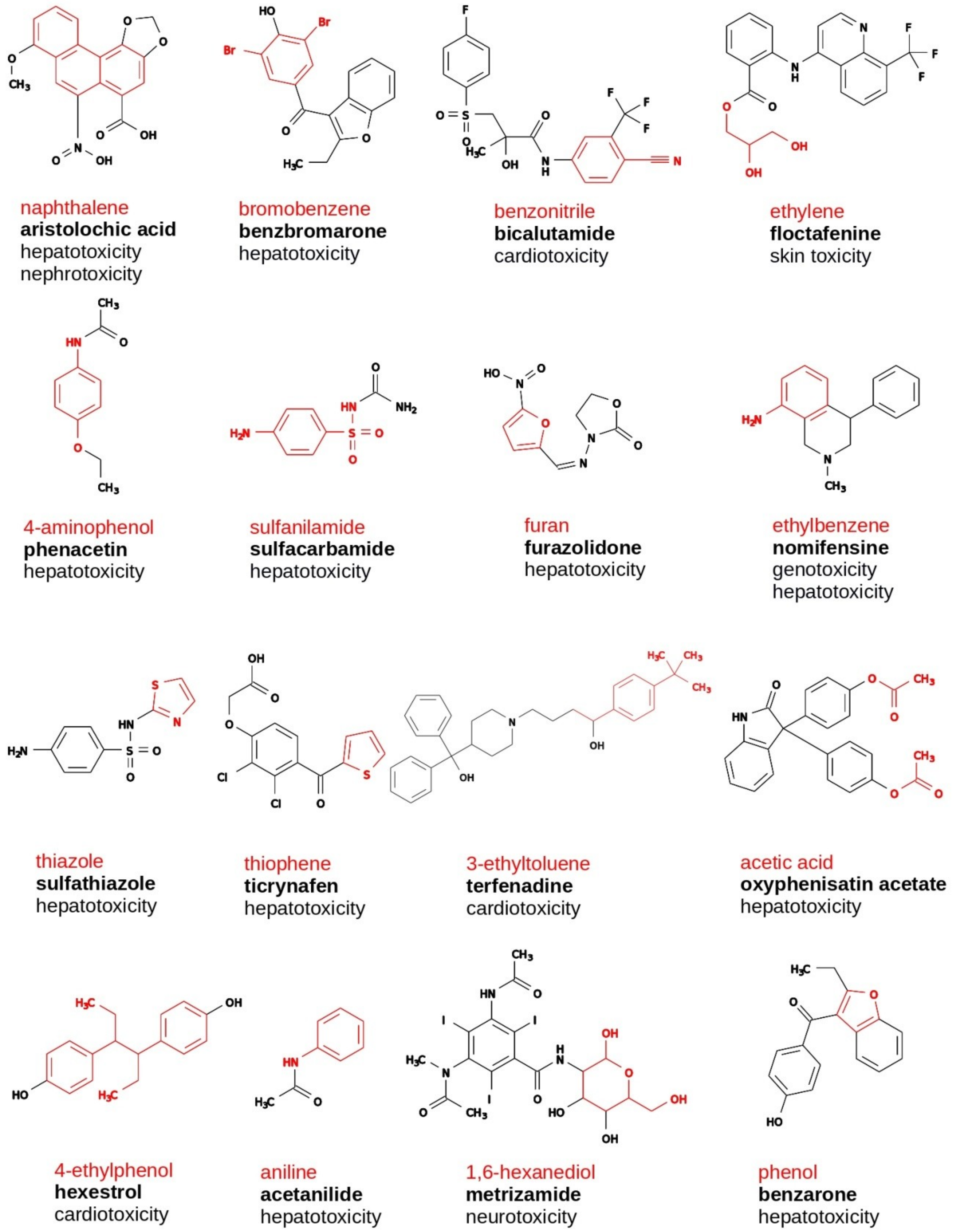


Figure 2.2 Examples of chemical fragments (highlighted in red) that predicted to lead to toxicities in drugs withdrawn from the market.

2.4.4 Prediction via sparse canonical correlation analysis

To explore how chemical fragments might be linked to toxicity endpoints, we employed a methodology similar to one used previously (Pauwels et al., 2011; Seoane et al., 2014), namely sparse canonical correlation analysis (SCCA). This technique can obtain the maximized correlation between the linear combinations of variables from two high-dimensional heterogeneous datasets, e.g., a linear combination of chemical fragments and a linear combination of toxicity endpoints. SCCA can display a global view of the dataset by reducing variables into only a small number of dimensions that capture the main properties, for instance, chemicals containing the same substructures could have the similar toxicity endpoints. In our simulated study, two datasets containing categorical variables (e.g., 0, 1) - X (compound-fragment matrix) and categorical variables (e.g., 0, 1) - Y (compound-toxicity matrix), consisting of p (415 fragments) and q (232 toxicity endpoints), respectively, and n = 991 samples were used. First, we normalized variables of X and Y by subtracting the column means and dividing by the column standard deviations to delete null columns where all values were the same. For simplicity, the simulation was standardized by replacing variance-covariance matrices with identity matrices in equal 2.3. Then K is computed as the covariance between standardized datasets X t and Y t:

$$K = \mathrm{Cov}(X\ t\ , Y\ t\ ) = \Sigma_{xt\ Y\ t} \qquad (2.5)$$

The results of this process are weight vectors, u1 and v1, in the linear combinations of variables from datasets X and Y.

2.4.4.1 Performance evaluation

We accessed the performance of SCCA using 5-fold cross-validation, thus datasets X and Y were split into five groups and each group takes turns to be the test set. **Figure 2.3** shows the ROC curve for the SCCA method where predictions for all toxicity endpoints are combined. With the parameters c1 = c2 = 0.04 and m = 100 for SCCA, the model obtained the best AUC score (0.801). The accuracy of individual toxicity endpoints (boxplot in **Figure 2.4**), obtained with the parameters c1 = c2 = 0.04 and m = 100.

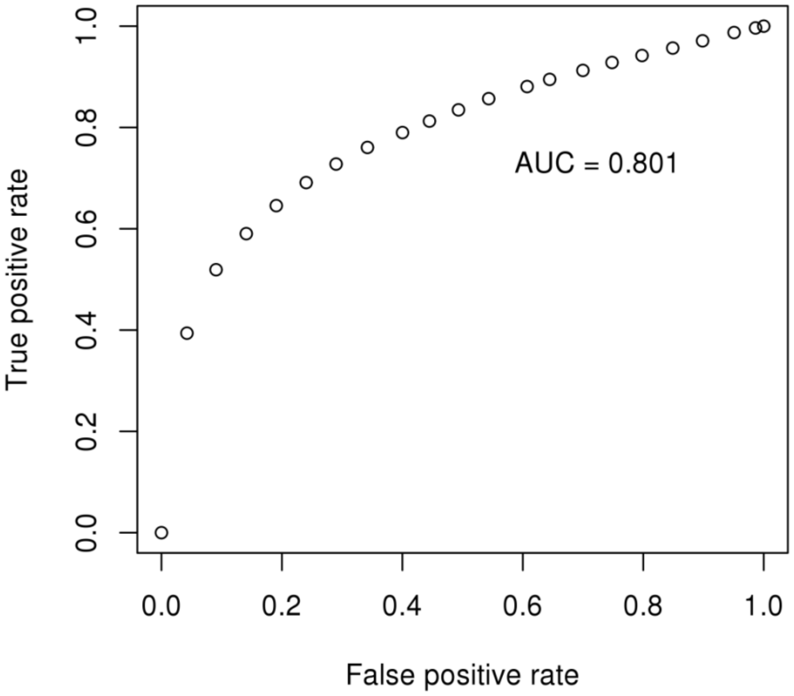


Figure 2.3 ROC curves in the 5-fold cross-validation to evaluate the performance of SCCA

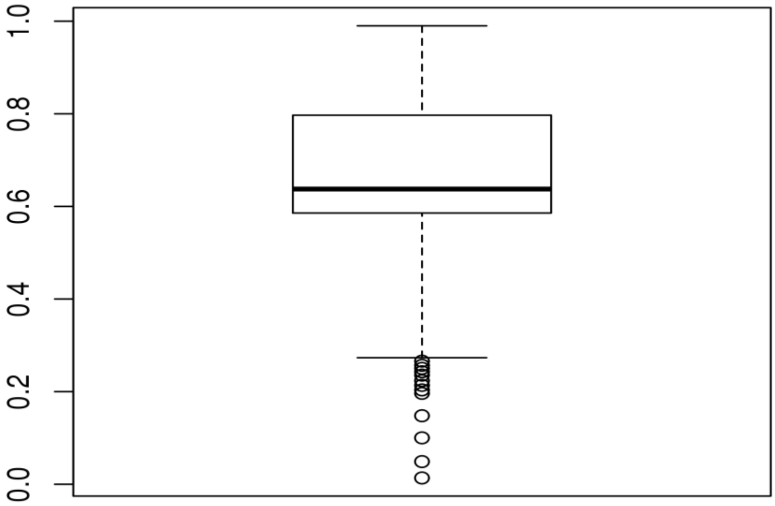


Figure 2.4 Boxplot of the AUC scores for individual toxicity endpoints to evaluate the performance of SCCA

The weight vectors for chemical fragments and toxicity endpoints were investigated and the index-plots of the first three canonical components were shown in **Figure 2.5**. Most all elements in the weight vectors in SCCA are zero in each component, which means SCCA can extract a small number of more selective and informative correlations between fragments and toxicities without losing performance.

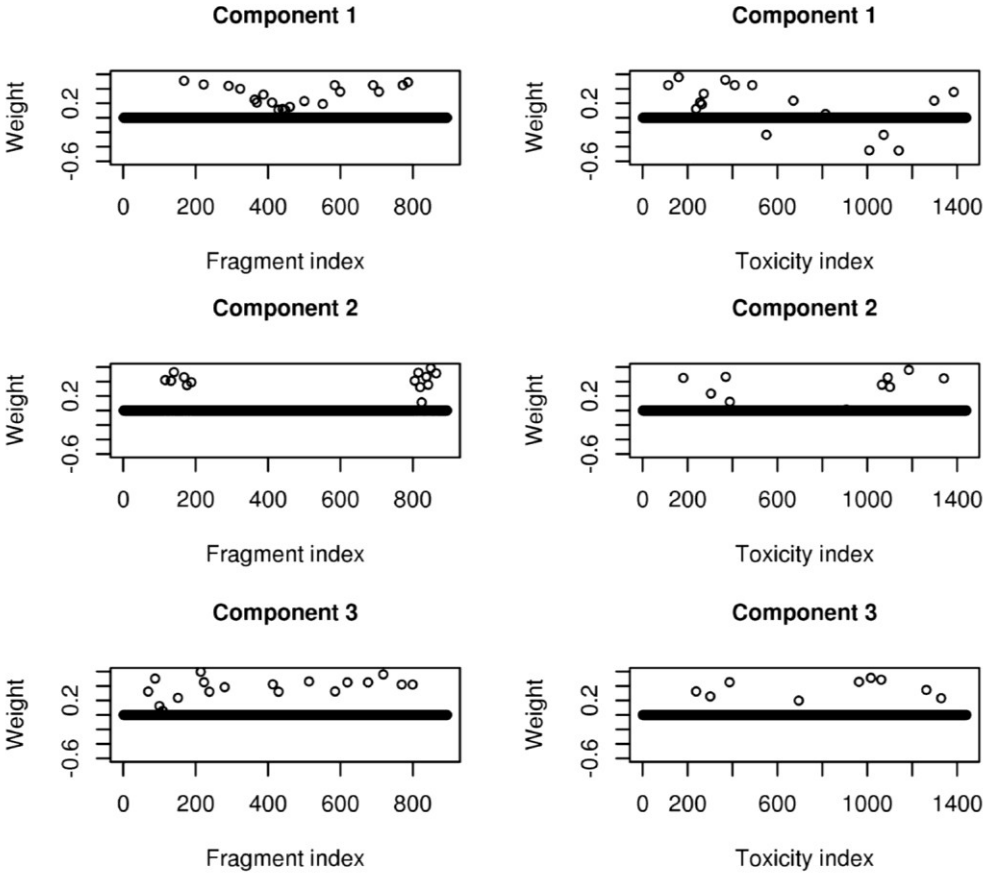
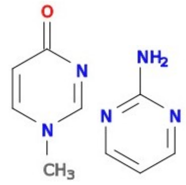


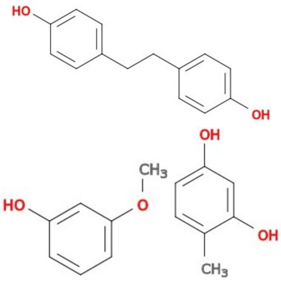
Figure 2.5 Index-plot of weight vectors for chemical fragments (left) and toxicity endpoints (right) extracted by SCCA

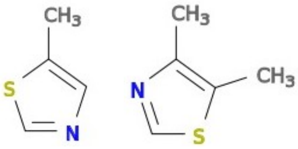
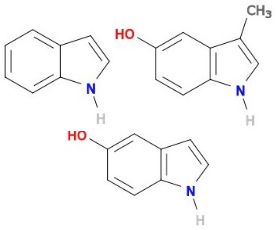
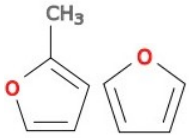
2.4.4.2 Extracted sets of chemical fragments and toxicity endpoints

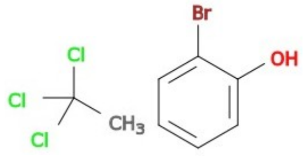
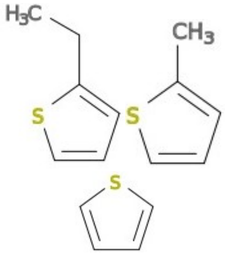
The above results prompted us to examine selected correlations of chemical fragments with toxicity endpoints in each canonical component extracted by SCCA for biological interpretations. Each component shows only a small set of chemical fragments related to a small set of toxicity endpoints. In each component of SCCA, there are two sets of compounds with high scores selected: one is the compounds with highest scores containing the associated chemical fragments, and the other is the compounds with highest scores containing the associated toxicity endpoints. A correlation coefficient is calculated to estimate the importance of the component correlation. **Table 2.4** reports some components with higher coefficients clustering specific chemical fragments and toxicity endpoints in a small number of compounds. The results for the complete list can be found in **Table S2.3**. Most putative mechanisms of toxicity for the fragments are unknown, however, some plausible mechanistic explanations could be predicted by interrogation of the literature.

Table 2.4 Chemical fragments with significant correlations to toxic endpoints

Canonical component	Correlation coefficient	Fragments with high scores	Toxicity endpoints with high scores	High scoring compounds (fragments)	High scoring compounds (toxicity endpoint)
59	0.988	 pyrimidine derivatives	glioblastoma necrosis	theophylline; tegafur; idoxuridine; floxuridine	theophylline

55	0.980	
		phenol derivatives
			thyroid fibrosis	equol;
				6,8-dimethoxy-3-methyl-3,4-dihydroisocoumarin;
				3-(2,3,4-trimethoxyphenyl)acrylic acid
					equol

48	0.816	 thiazole derivatives	heart muscle necrosis	thiamine; pramipexole	pramipexole
18	0.777	 indole derivatives	liver ischaemia	melatonin; hydroxyindoleacetic acid; 5-hydroxytryptophan; norharman; harmine; clausine E	clausine E; serotonin; melatonin; hydroxyindoleacetic acid; 5-hydroxytryptophan
65	0.767		t-cell necrosis	urinastatin; furapyrimidone; 5-nitro-2-furylpropionamide; 4-hydroxynitrofurazon	furapyrimidone; 4-hydroxynitrofurazone

				e; methoxsalen; furanodienone; 2,3,4,7,8-pentachlorodibenzofuran	
75	0.706	 halogenated derivatives	testis degeneration	pentabrominated diphenyl ether; DDT	pentabrominated diphenyl ether; firemaster BP-6; dichlorodiphenyl dichloroethylene; DDT
36	0.701	 thiophene derivatives	coagulation	ticlopidine; methapyrilene	ticlopidine; fenfluramine; aspirin

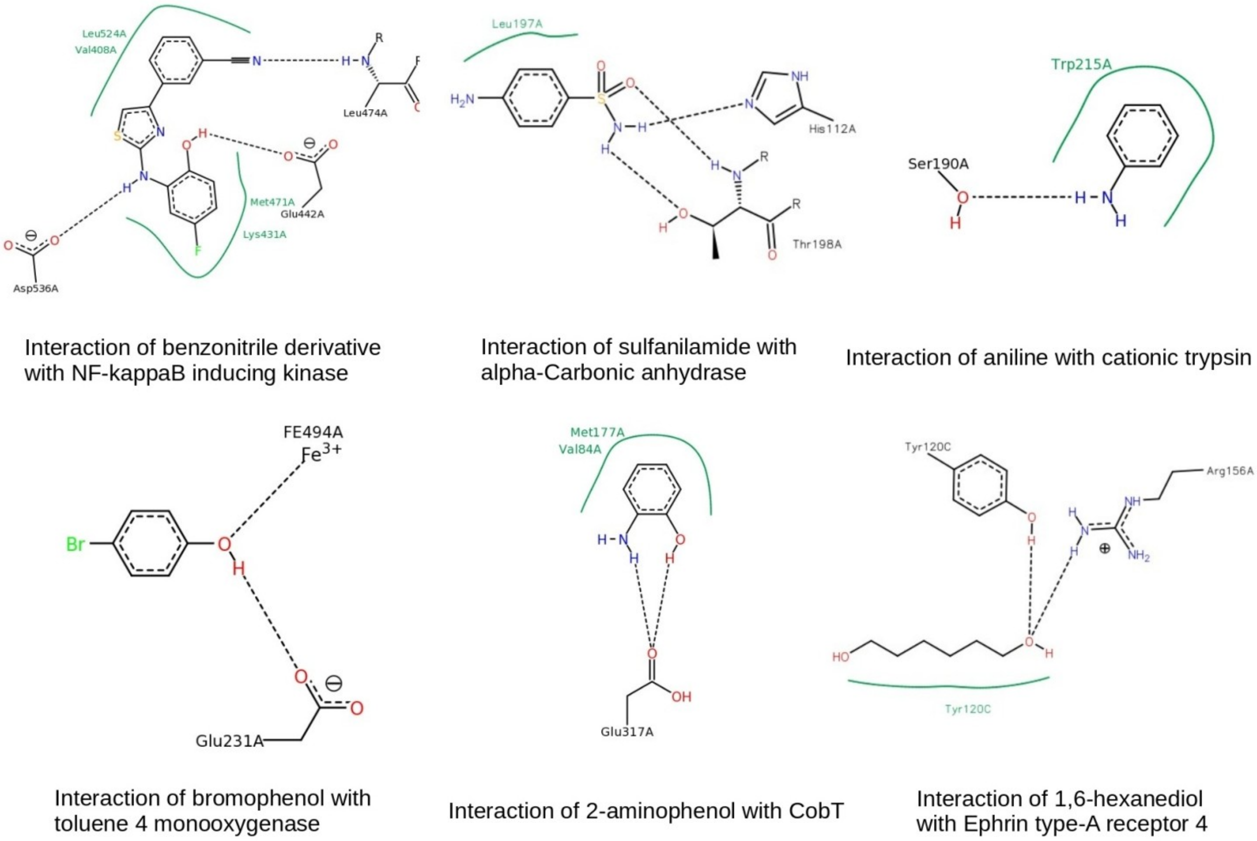
Compared with frequency analysis, SCCA extracted associations of a set of fragments with similar structures and a set of toxicity endpoints. The substructures responsible for toxicities can be obtained according to chemical categories: indole derivatives, thiophene derivatives, thiazole derivatives, phenol derivatives, etc. This makes the explanations of mechanisms more straightforward. For example, component 75 contains the halobenzene and halogenated hydrocarbon substructures linked to testis degeneration. The toxic effects are likely from cytochrome P450-dependent reduction to the reactive forms (halobenzene epoxide and trichloromethyl radical) which can bind to proteins to induce the toxicity (Allen et al., 1979; McGregor and Lang, 2000).

2.4.4.2 Toxicity prediction for withdrawn drugs

The prediction results for withdrawn drugs are shown in **Table S2.4**. Compared with SVM, SCCA tends to predict the specific toxicities (e.g., pulmonary oedema, neutrophil phagocytosis) that are associated with special drug clusters, while the common toxic terms (e.g., hepatotoxicity, nephrotoxicity) are rarely predicted with high scores for a drug. For example, amphetamine, dexfenfluramine, tranylcypromine, terfenadine were predicted with higher scores for a specific toxic term “valvulopathy” - a main cause of congestive heart failure (Carabello and Crawford, 1997), of which amphetamine and dexfenfluramine were reported to associate with cardiac valvulopathy (Rothman et al., 2000).

2.4.5 Mechanism study of toxicity

Ideally, statistically significant predictions should be complemented with putative mechanistic explanations for why they are predicted. An emerging theme from the literature about several fragments suggests the finding that covalent binding by structural alert moieties to cellular proteins could be a direct cause of toxicities. Several of these moieties have been seen within protein three-dimensional structures. We identified these by searching for their occurrence in ligands within known structures (**Figure 2.6**).



Interaction of benzonitrile derivative with NF-kappaB inducing kinase

Interaction of sulfanilamide with alpha-Carbonic anhydrase

Interaction of aniline with cationic trypsin

Interaction of bromophenol with toluene 4 monooxygenase

Interaction of 2-aminophenol with CobT

Interaction of 1,6-hexanediol with Ephrin type-A receptor 4

Figure 2.6 Interactions of selected structural alerts with biomolecules (black dashed lines show hydrogen bonds and green lines indicate hydrophobic interactions). The interactions provided by PoseView are extracted from the protein databank (PDB; rcsb.org).

It is also important to note that toxicity owing to a structural alert is conditional. The toxic structural alert-containing drugs could be hazardous or safe depending on their metabolic pathways and the bioactivity of the structural alert (Rybacka et al., 2015). Most toxic cases are caused by the reactive metabolites of the drugs instead of the drugs themselves.

XenoSite can predict when structural alerts of furan, phenol, nitroaromatic, and thiophene will be bioactivated in drugs (Dang et al., 2017). We used the model of SMARTCyp in XenoSite to predict the six thiophene-containing drugs: methapyrilene, suprofen, ticrynafen, zileuton, dorzolamide, and olanzapine (Rydberg et al., 2010) (**Figure 2.7**). The first four drugs are shown to form reactive metabolites through the S-oxidation or epoxidation of thiophene (colored in red) by CYPs and they have been withdrawn from the market, while the thiophene groups contained by the last two drugs dorzolamide and olanzapine do not undergo bioactivation nor toxic (colored in green and blue).

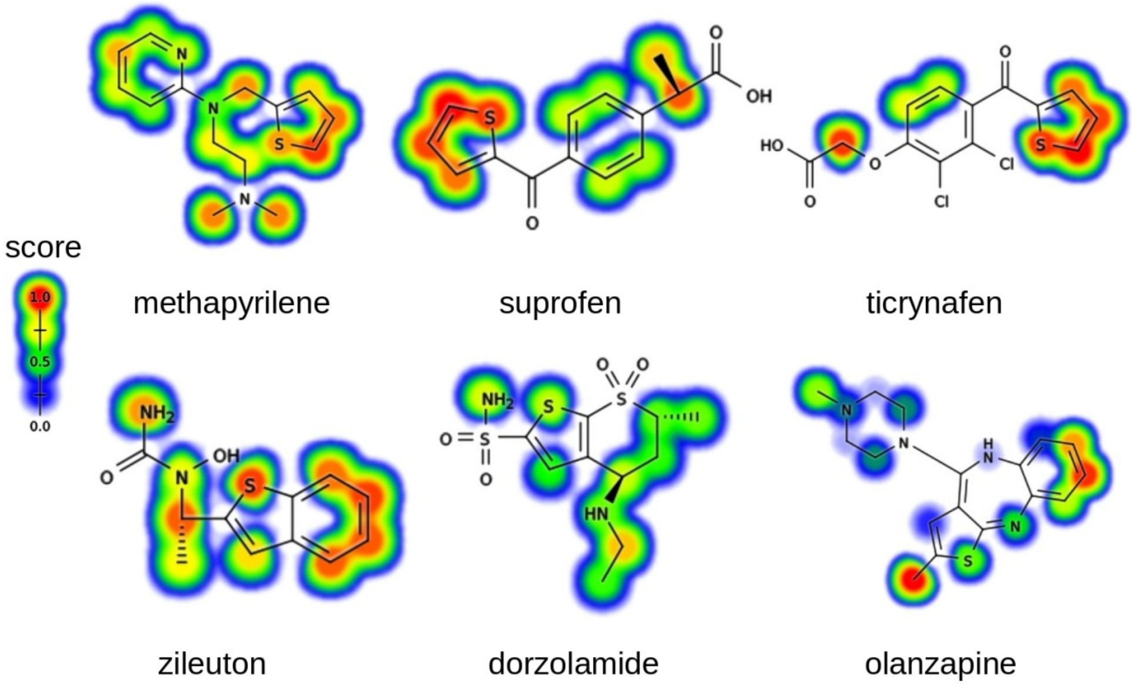


Figure 2.7 Metabolism model predicts whether thiophenes are bioactivated by CYPs. SMARTCyp model is used to shade atoms of drugs, ranging from white (0.0) to green (0.5) to red (1.0). The red represents the bioactivated parts and the blue parts are not bioactivated.

Besides direct drug targets such as nuclear receptors, drug transporters, drug metabolizing enzymes with known toxicity associations, other protein interactions (i.e. off-target interactions) might also contribute to toxicities. We constructed a network (**Figure 2.8**) by integrating the interactions of the structural alert benzbromarone with gene expression data from TG-GATEs and associated pathways defined by gene ontology (GO) categories (Ashburner et al., 2000). A more systematic study relating gene expression to toxic response is given in Chapter III. Benzbromarone was withdrawn from the market in France due to its hepatotoxicity. It is metabolized by CYP3A4 to 1'-

hydroxybenzbromarone and by CYP2C9 to 6-hydroxybenzbromarone, which is be further metabolized by CYP2C9 and CYP1A2 to 5,6-dihydroxybenzbromarone. The structural alert Acyclic di-aryl ketone moiety can be oxidized to 2,6-dibromohydroquinone and 2,6-dibromobenzoquinone which might induce hepatotoxicity by covalent binding with proteins (Kitagawara et al., 2015).

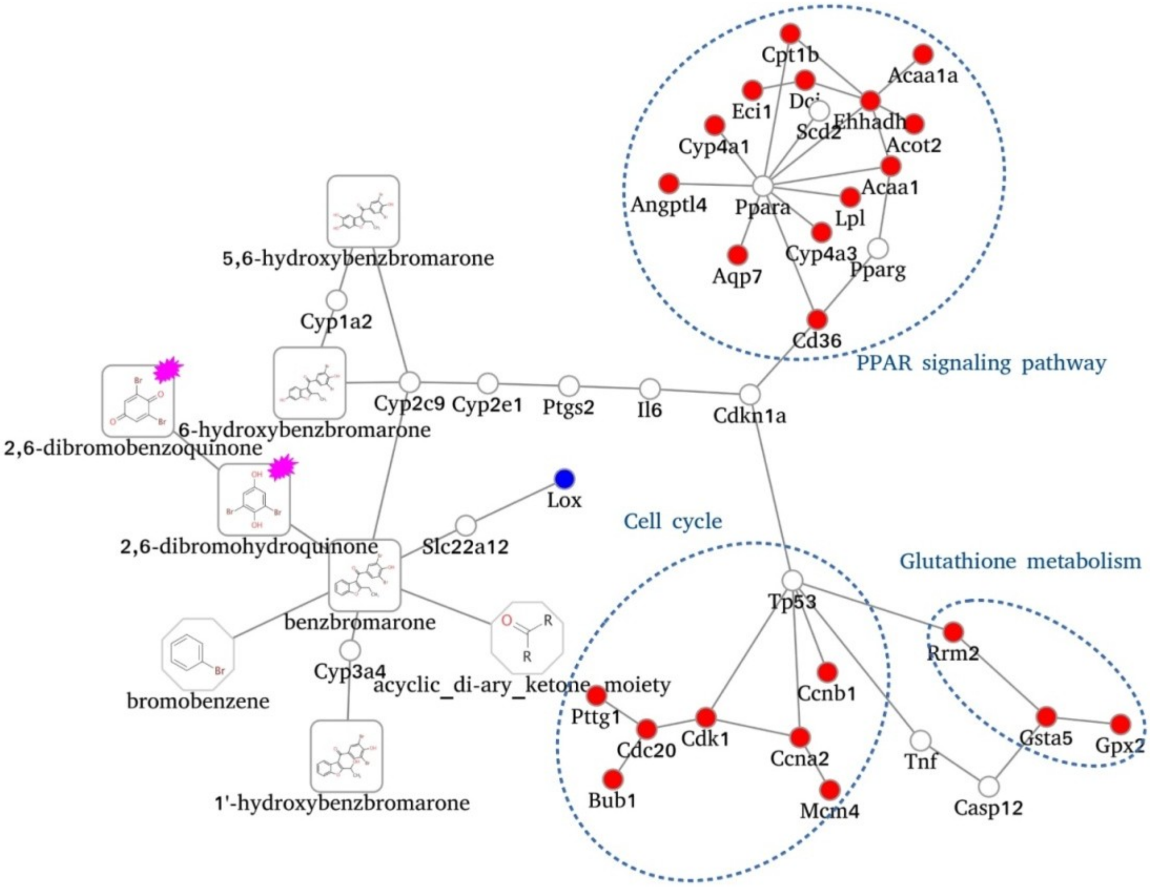


Figure 2.8 Network of benzbromarone-induced hepatotoxicity response. Genes are shown in circles; those in red are up-regulated and blue are down-regulated in rat liver (*in vivo*). Structural alerts are highlighted with stars.

Single nucleotide polymorphisms in CYP2C9 might be associated with the individual toxicity response to benzbromarone (Takahashi and Echizen, 2003). 36 variations of CYP2C9 from 1000 Genomes Phase 3 (The 1000 Genomes Project Consortium, 2015) were tested using Mechismo (Betts et al., 2015). Four of these, R124L, S127F, M240T and I359L at the binding sites are predicted to affect the bindings of triflucan, bifonazole, flurbiprofen and piperazine, which might increase and/or inhibit the effectiveness of those drugs (**Figure 2.9**). A systematic analysis of associations between polymorphisms and toxicity is discussed in Chapter V.

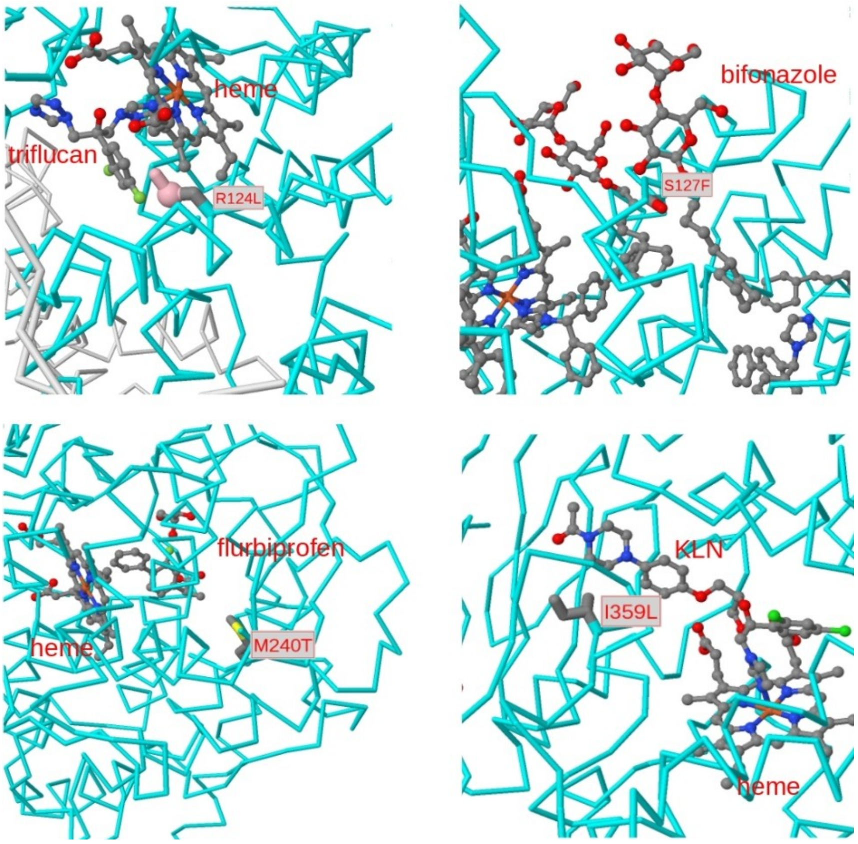


Figure 2.9 Structures of CYP2C9 mutations in contact with four drugs

2.5 Conclusion and discussion

The systematic analysis of a large number of chemical compounds and their links to toxicity terms revealed that statistics (e.g., frequency analysis) and machine learning methods such as both SVM and SCCA are capable of identifying structural alerts associated with toxicity. We found hundreds of structural alerts, including some known alerts like aniline, furan, or phenol. Many of the identified fragments are known to form reactive metabolites that might induce toxic effects. This finding has been used previously to aid quantitative structure-toxicity relationships (QSTR) analyses (e.g., quetiapine versus clozapine (Nelson, 2001)), suggesting that the presence of the fragments leads to a higher incidence of toxicity and that they should be avoided in drug development where possible.

Nevertheless, some structural alerts can be beneficial for activity. For example, quetiapine and clozapine's structurally similar agent, olanzapine, contains the aniline structural alert and is indeed known to form reactive metabolites, and associated covalent bonds with proteins, but is nevertheless devoid of toxic effects (Nelson, 2001). Predictions from such alerts thus need to be considered with caution.

The results in this chapter also hint that toxicity predictions should be coupled to detailed mechanistic analyses. A comprehensive understanding of how bioactivation pathways produce toxic effects can come by integrating gene expression data from *in vivo* and *in vitro* models. Other studies suggest this strategy could be fruitful. For example, the withdrawn antibiotic trovafloxacin shows significantly different gene expression patterns (related to genes involved in oxidative stress response) compared to structurally similar fluoroquinolones still on the market (Stepan et al., 2011). A recent study also suggests that shifts in IL-6/TNF-α ratios play a role in immune-mediated trovafloxacin-induced hepatotoxicity (Bonzo et al., 2015).

Several genetic studies indicate that polymorphisms in drug metabolizing enzymes could lead to individual differences in reactive metabolite formation and thus different responses to xenobiotics (Williams and Park, 2003). For instance, it has been suggested that polymorphisms in the enzyme N-acetyltransferase 2 (NAT2) are responsible for the severity of hepatotoxicity of isoniazid, an antibiotic used for the treatment of tuberculosis. The mutants of NAT2 were found to increase the risk of hepatotoxicity by activating the hydrazine group (a structural alert) to form N-acetylisoniazid intermediate (Huang, 2007).

Overall, application of structural alerts to filter drug candidates should be complemented by putative mechanistic insights based on existing biological data. This is a subject that we discuss in the following chapters.

CHAPTER III: Deciphering mechanisms of drug action and toxicity by integrating gene expression signatures, biological features and chemical features

3.1 Abstract

The molecular mechanism for how xenobiotics exert beneficial or detrimental effects on biological systems is important for many applications in pharmacology and biomedicine. Gene expression screens performed systematically on many compounds in multiple tissue/species systems provide opportunities to investigate principles of the molecular impact of xenobiotics on biological systems. We analyzed the TG-GATES gene expression and pathology data on 131 compounds in liver or kidney tissues and cell lines. We found that *in vivo* and *in vitro* data are only rarely similar and gene expression profiles of *in vivo* rat liver data are more informative about mechanism than other datasets. Clustering compounds according to expression signatures forms groups with similar therapeutic characteristics and/or mechanism of drug actions, with relatively few clusters arising owing to a similarity in toxic response, suggesting that drug mechanism-of-action is the predominant effect observed in expression studies. We also found correlations between overexpression of genes related to cell proliferation and drug metabolism with compound solubility, and between structure alerts (e.g. acetanilide, 4-chlorotoluene) and the expression of genes associated with the mechanisms of drug-action, although we found no correlation between any measure of gene expression and molecular structure similarity. Our findings suggest that simultaneous exploration of biological, chemical features and gene expression changes can enrich understanding of drug action and ultimately help drive alternatives to animal models.

3.2 Introduction

The advent of high-throughput techniques to study gene expression led, when used in toxicology studies, to the field of toxicogenomics, where changes in gene expression are used to identify potential markers of compound toxicity (Afshari et al., 2011). There have been a number of success stories, for example, where particular biomarkers or gene signatures have been identified using these techniques (Caiment et al., 2014; Ichimura et al., 1998; Yamada et al., 2012). Moreover, despite many studies both in toxicogenomics, but also more generally in chemically induced gene expression changes, the relationship between a chemical features and biological response remains elusive. Gene expression profiling has been a popular technique to study responses to cell or tissue stimuli and has been used, for example, to detect differences between normal and diseased tissues (Velculescu et al., 1999), to detect RNA expression responses in tissues that have been treated with drugs (Gray et al., 1998) and to detect the mechanisms underlying biological pathways (Bertilsson et al., 1998). Previous studies have shown that chemical features can be related to gene expression. For example, there are correlations between chemical structure of cancer drugs and gene expression (Blower et al., 2002). Elsewhere distinct chemical properties correlate with specific cellular responses: for instance, hydrophobic properties correlate strongly with DNA damage response, and hydrogen bonding is linked to metabolic stress (Khan et al., 2012). A reliable connection between one class of up-regulated genes induced by certain compounds in the liver and their chemical information found in ECFP4 fingerprints has also been predicted (Fernald and Altman, 2013).

More recently, a variety of public and commercial databases of expression profiles have been developed and presented many advantages to the identification of pharmacological or toxic phenotype of new chemical entities. For example, the Connectivity Map dataset contains (in build 02) more than 7 000 microarray expression profiles from cultured human cells treated with 1 309 bioactive small molecules and has been used in drug repositioning and for elucidating the mechanism of action of drugs (Lamb et al., 2006). The Open Japanese Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATEs) consists of gene expression data and pathological data from 131 different compounds, mainly medical drugs, given *in vivo* and *in vitro* to rats and human at different doses and time points and then measured in liver and kidney tissues and cell-lines (Uehara et al., 2010). These data have been employed in studies to identify candidate biomarker genes and to construct prediction models for hepatotoxicity (Gao et al., 2010; Hirode et al., 2009; Kiyosawa et al., 2007; Omura et al., 2007; Uehara et al., 2011, 2008a), nephrotoxicity (Kondo et al., 2009) and hepatocarcinogenicity (Uehara et al., 2011, 2008b). Such studies have provided data and approaches to systematically compare the gene expression profiles *in vivo* and *in vitro*, across different species and within different tissues, and to explore the relationships between biological features, chemical features, gene expression profiling and pathology. In this chapter, we explore how toxicogenomics signatures (within the TG-GATEs project) relate to chemical features and how well expression changes capture toxic or other mechanistic information related to drug response.

3.3 Methods

3.3.1 Toxicogenomics datasets

The TG-GATEs (http://toxico.nibiohn.go.jp/english/) database (Phase I) contains 131 compounds, mainly drugs (Uehara et al., 2010). For the *in vivo* studies, selected rats were treated with both a single-dose, consisting of four time points (3h, 6h, 9h and 24h) with three doses (low, middle and high), and a repeated-dose, consisting of four treatment durations (4d, 8d, 15d and 29d) with three dose levels (low, middle and high). Gene expression of livers and kidneys from the rats were profiled with three animals in each group. In the *in vitro* studies, the primary rat hepatocytes and primary human hepatocytes were treated at three time points (2h, 8h and 24h) with three concentrations (low, middle and high) by each of the 131 compounds. All the studies had time-match controls.

3.3.2 Microarray data and pathological data processing

Affymetrix CEL-files were downloaded from the Open TG-GATEs database (http://toxico.nibiohn.go.jp/datalist.html) and processed by the Bioconductor package *simpleaffy* (Wilson and Miller, 2005) for generation of expression values for each probe set using the robust multiarray averaging (RMA) method (Irizarry et al., 2003). Probe sets were then reannotated to genes using custom CDF file (version 15.1.0) (Dai et al., 2005). Finally, the replicate samples were summarized by taking the average and differential expression was calculated as log 2-ratio compared to the time-matched controls. Statistical significance p-values were assigned using a modified t-test (Storey and Tibshirani, 2003). For subsequent analysis we considered genes which had a significant (at least 2 fold) increase or decrease ($p < 0.01$) in expression by any of the 131 compounds at any dose or time point (**Table S3.1**).

Histopathological records for the rat *in vivo* data were obtained from the TG-GATEs web site. Pathological observations were coded from 0 to 4 based on intensity by pathologists: 0=no change, 1=minimal, 2=slight, 3=moderate, 4=severe and values from replicate samples belonging to each instance were averaged. We then summed up the values of all histopathology findings in each organ and used them as an overall measurement of pathology classes (**Table S3.3**).

3.3.3 Principal component analysis (PCA) and hierarchical clustering analysis (HCA)

Principal component analysis (PCA) was performed to visualize the gene expression data in two dimensions, i.e., the first two principal components with highest variance. Specifically, a matrix (100 top-ranking genes vs compounds) was built with each element representing the expression of the 100 genes with highest fold change (absolute values) in *in vitro* rat hepatocyte (**Table S3.2**). Then, a PCA using the *prcomp* command in R (R Development Core Team, 2012)) was implemented on this matrix to investigate the batch effects in the TG-GATEs.

Furthermore, for investigating the gene expression similarity between any two compounds, we calculated a Jaccard index (intersection divided by union) as a measure of similarity, counting each gene/expression-direction pair once. We then used the hierarchical cluster analysis program, OC (GJ Barton, University of Dundee) to obtain compound clusters based on these scores. We employed “complete” linkage options to divide the whole set into discrete clusters. Tree figures were generated using iTOL (Letunic and Bork, 2011).

3.3.4 Biological data collection and visualization

We extracted ATC-classes (i.e. descriptions of therapeutic use) for all 131 chemical compounds from the WHOCC database (http://www.whocc.no/). Annotations of toxic endpoints induced by chemicals were extracted from the publications of toxicogenomics experiments (http://www.toxwiz.com/) and drug-target interactions were assigned using DrugBank (Knox et al., 2011).

We constructed networks of genes and their interactions from the rat metabolic and signaling pathways in KEGG (Kanehisa et al., 2012), and rat protein-protein interactions from STRING (Franceschini et al., 2013) and rat chemical-protein interactions from STITCH (Kuhn et al., 2014). We considered both experimental and predicted with high confidence (score > 0.7) interactions in STRING and STITCH.

3.3.5 Chemical data collection and visualization

We extracted structure files for 98 compounds from PubChem (Bolton et al., 2008). After adding hydrogen atoms, transforming to 3D structures, chemical properties (e.g. number of hydrogen bond acceptors and donors, LogP, topological polar surface area) were computed by Open Babel (v. 2.3.1) (O'Boyle et al., 2011). The structure similarity between any two compounds was calculated based on Molprint 2D using Tanimoto coefficient (Bender et al., 2004), as shown in Chapter II Methods.

We downloaded 162 261 clean fragment structures filtered according to the rule-of-three selection criteria from ZINC database (version 12, Irwin and Shoichet, 2005) and cleaned charges manually. For substructure searching, we used structural fingerprints from Open Babel (v. 2.3.1) and PubChem (Bolton et al., 2008).

We matched 1 332 distinct small fragments from ZINC database to 98 compounds that modulated gene expression *in vivo* rat liver. We then created a binary feature matrix with 98 rows (one for each compound), and 1 332 columns (one for each small fragment). Entries in the binary feature matrix were set to 1 if the fragment was found in the compound, otherwise the entries were 0. We then performed principal component analysis on this 98 x 1 332 matrix.

3.3.6 Sparse canonical correlation analysis (SCCA)

See sparse canonical correlation analysis (SCCA) in Chapter II Methods.

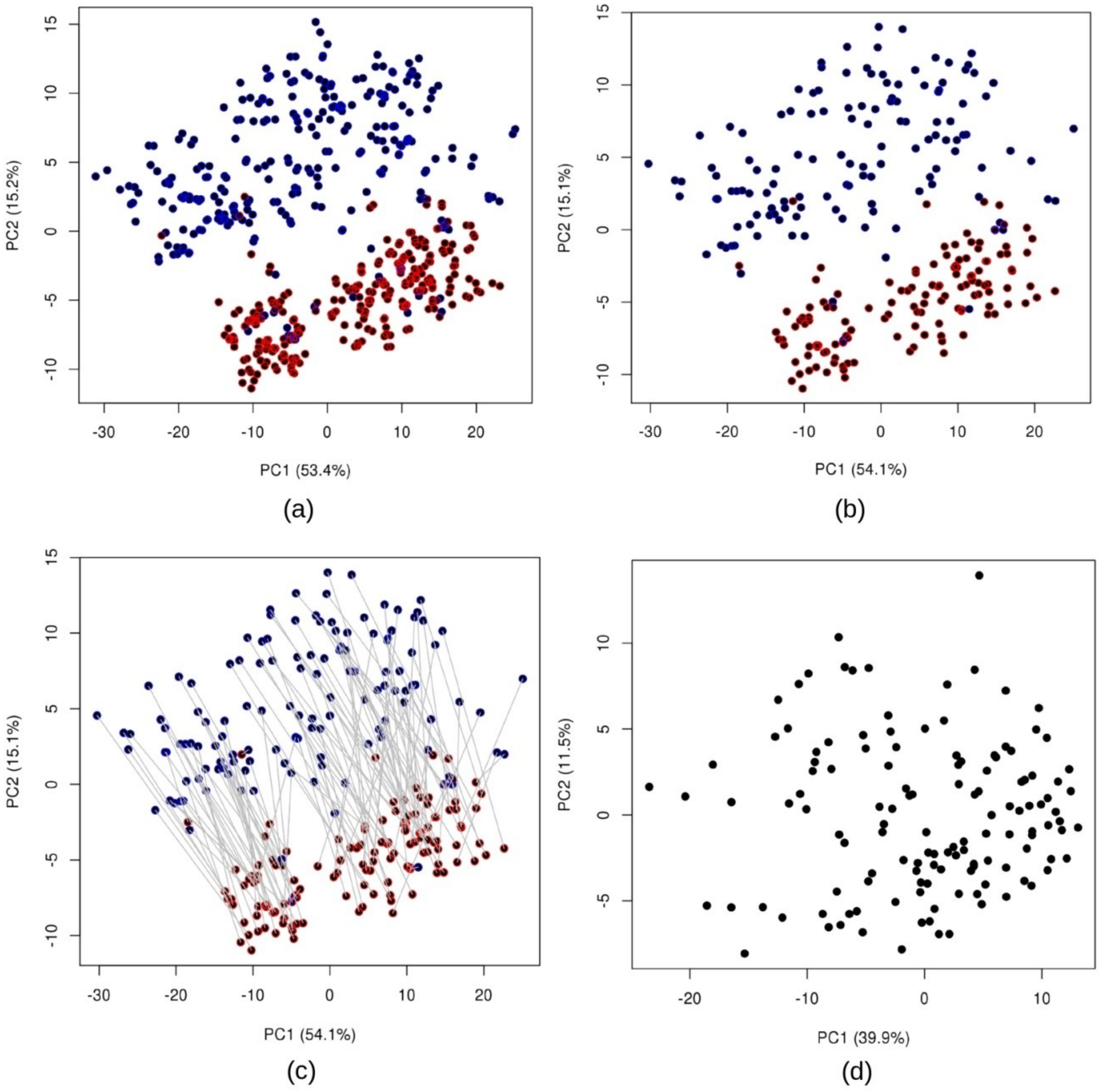
3.4 Results

3.4.1 Gene expression quality controls

Since the large scale of transcriptomics data were from several contributors, differences arising from subsets of sample batches are unavoidable. In order to identify the reliability of the genes and eliminate implausible data, here, we executed curation steps including controlling batch effects, assessing reproducibility across replicates and analyzing concentration-response relationship presented by Grinberg et al. (Grinberg et al., 2014).

3.4.1.1 Batch effects

PCA was performed for the expression values of 229 probe sets representing the 100 genes with the highest fold change (absolute values) across 130 compounds (carbon tetrachloride was excluded due to unavailable data) in the 24h-high concentration of rat *in vitro* subset (**Table S3.2**). PCA applied to two replicates and two controls for each probe which illustrates their locations within four clusters (**Figure 3.1 (a)**). Using the mean values of the replicates and the mean values of controls gives similar clusters (**Figure 3.1 (b)**). When controls and the corresponding treated samples are connected by lines, we see two main clusters revealing that the difference between the clusters is a consequence of batch effects in microarray data (**Figure 3.1 (c)**) and that, after subtracting the controls from the corresponding treated samples, the initial clusters are not observed any more (**Figure 3.1 (d)**), suggesting that batch effects were eliminated by this step.



(a)

(b)

(c)

(d)

Figure 3.1 Principle component analysis for rat *in vitro* gene expression data obtained after treatments of 131 compounds with high dose level at 24h time point. The red and blue symbols represent the controls and treated samples, separately **(a)** PCA for all controls and samples **(b)** PCA for using the mean values of controls and replicates **(c)** Connecting lines between controls and the corresponding treated samples **(d)** PCA

after subtracting the controls from the corresponding treated samples

3.4.1.2 Reproducibility

We checked the distance between replicates and compared this to the influence caused by the test compound. **Figure 3.2** shows the frequency distributions for Euclidean distances between replicates and control-treatment pairs tested at the 24h time point at high concentrations, where 90% of distances between replicates are lower than 14, while 92.5% distance between control-treatment pairs are larger than 14. The median distance between replicates is 5.3-fold lower than that between control-treatment pairs, which suggests that the reproducibility between replicates is acceptable.

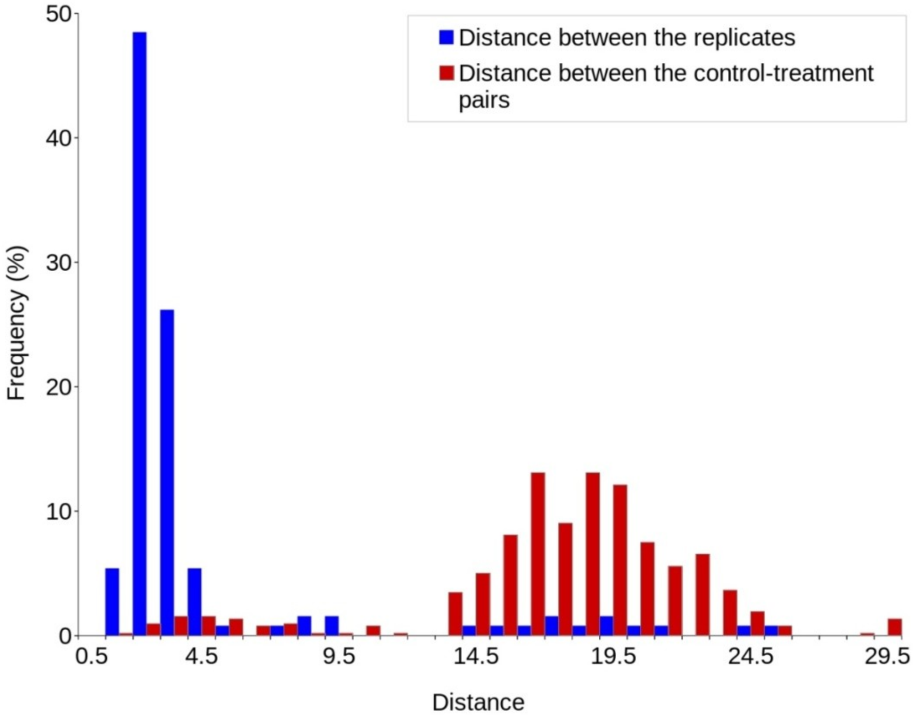


Figure 3.2 Frequency distribution of Euclidean distance between all replicate sample pairs (blue) and control-treatment sample pairs (red) for 24h high dose treatment

3.4.1.3 Number of dysregulated genes per compound

The number of dysregulated genes among the 100 top-ranking genes (**Table S3.4**) in rat *in vitro* subset was checked separately for all nine combinations of dose level (low, middle, and high) and time-point (2h, 8h, and 24h) (**Figure 3.3**). It shows that the number of dysregulated genes is different between the compounds and the top 100 genes were dysregulated by 112 of 131 compounds. Some of the well-known toxic compounds such as carbon tetrachloride weakly dysregulated a few genes in TG-GATEs. In the comparisons of the high versus middle concentration for the exposure periods of 2h, 8h, and 24h, giving number of compounds ratios of fewer regulated to more regulated genes of 1/20, 1/77, and 1/95, respectively, showing overall greater changes in response to a higher dose. These ratios are 3/3, 3/23, and 2/42 when comparing the middle to low concentration, also showing generally a stronger effect for a higher dose. The data in which the genes dysregulated at a low concentration remain unaltered at a higher concentration should be treated with caution.

3.4.1.4 Concentration-response relationship

We then investigated the concentration-response relationship for individual genes (**Table S3.5**). All upregulated genes with at least 2-fold change at any dose level (low, middle, or high) after 24h treatment with 131 compounds were determined (**Figure 3.4**). Of the 42 compounds which induced significant upregulation of a minimum of ten genes, 16 show that genes become upregulated from the low to the middle and from the middle to the high concentrations (e.g., tannic acid, clofibrate, naproxen and colchicine). For 22 compounds we see differences between high and middle, but not middle and low (e.g., cephalothin, paraverine, ibuprofen and sulpiride). Two compounds (phenobarbital and acetaminophen) upregulated most genes from the low to the middle but downregulated from the middle to the high concentration. For two others (puromycin aminonucleotide and WY-14643), genes are upregulated at low, but not at middle or high doses. There is no obvious explanation for these exceptions, though it is possible that higher doses lead to drastic alterations in physiology (possibly leading to death of the animal) that are not captured in the expression set. Nevertheless, the majority of compounds show an expected drug-concentration-response relationship.

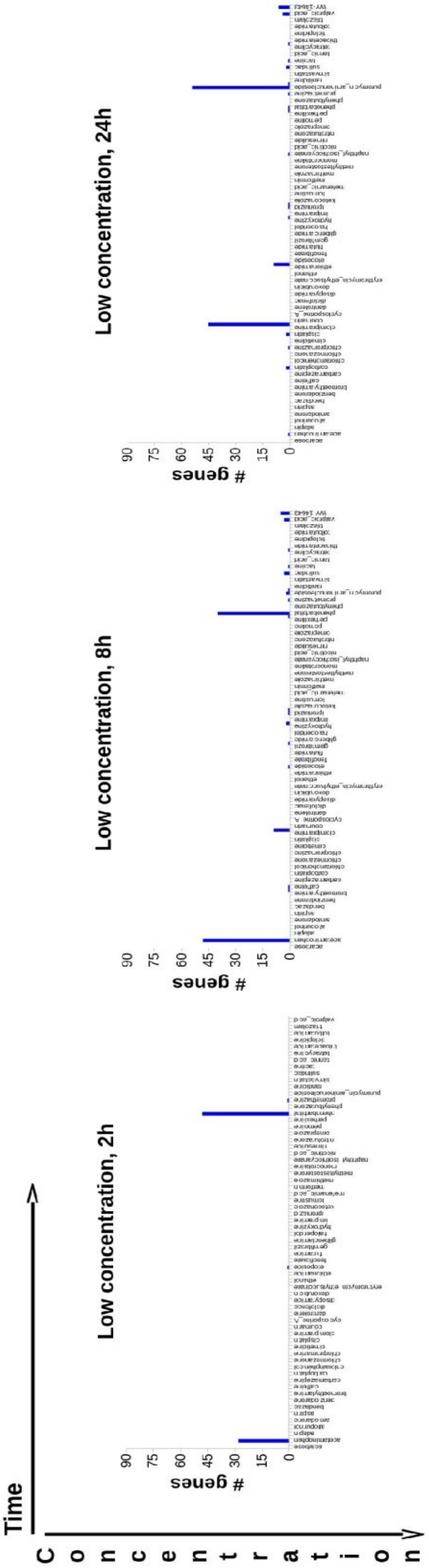
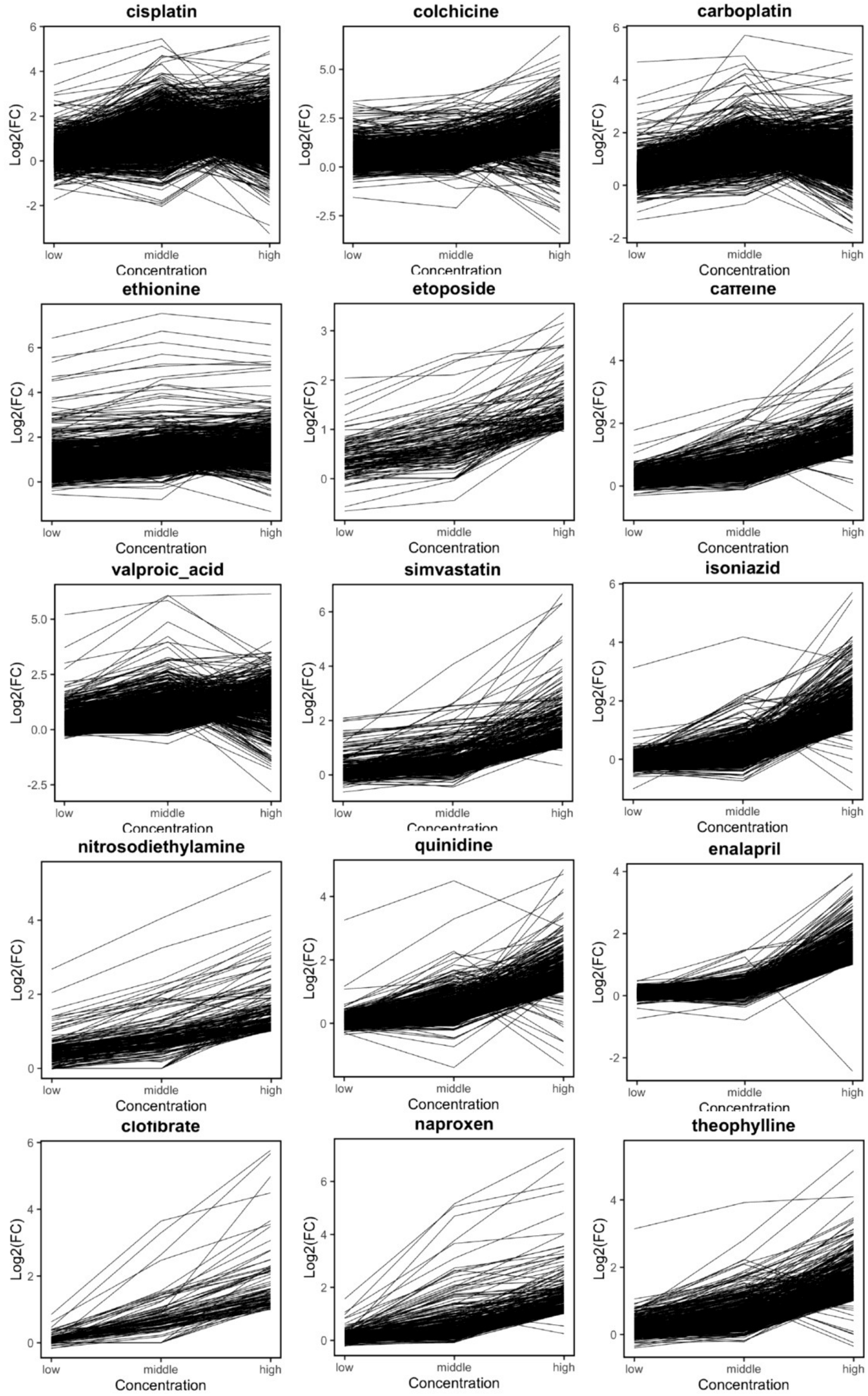
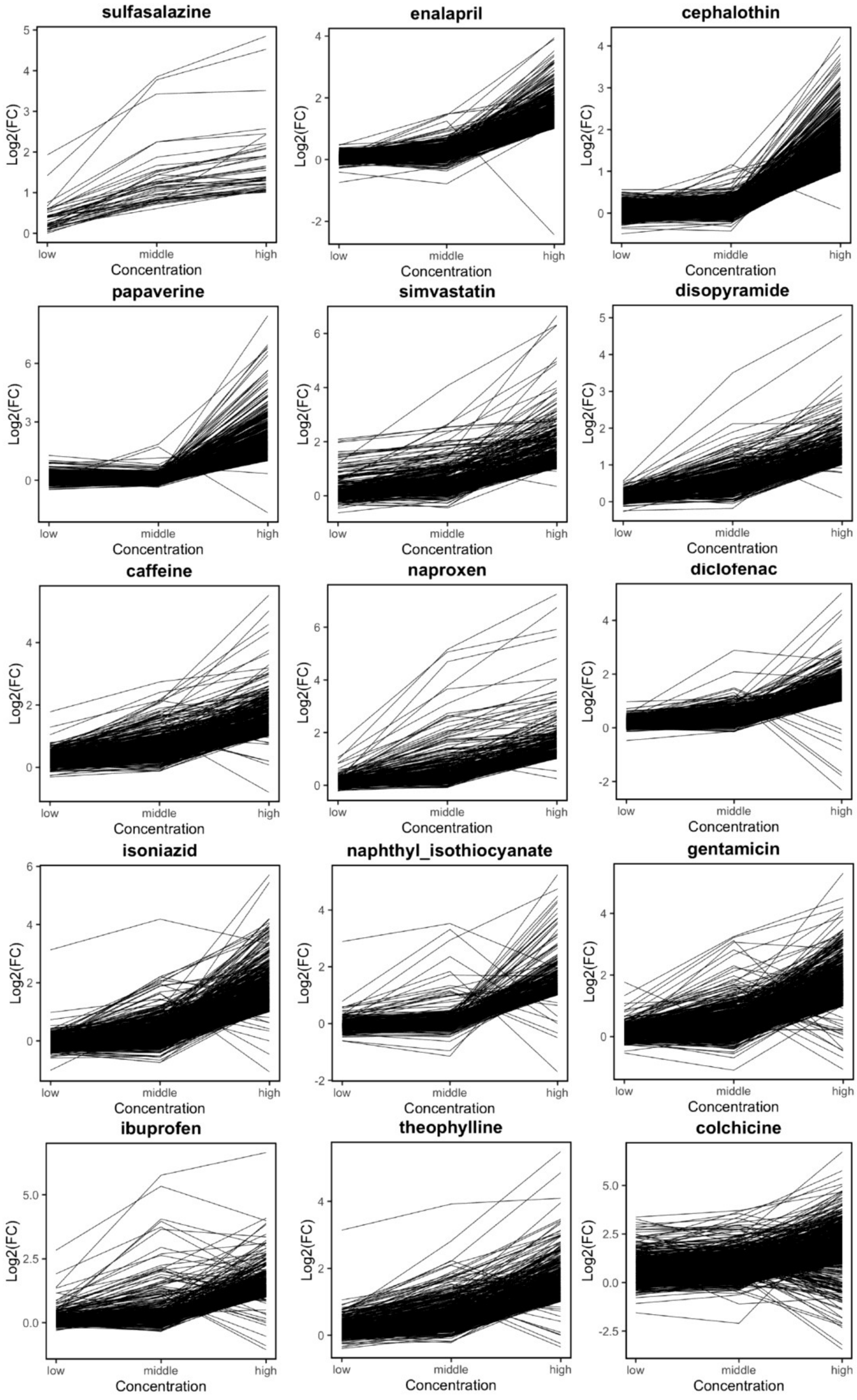


Figure 3.3 Number of dysregulated genes per compounds. The X axes lists 112 compounds that were tested at the indicated time point and concentration and the Y axes illustrates the number of genes affected among the 100 strongest dysregulated genes.


cisplatin
colchicine
carboplatin
ethionine
etoposide
caffeine
valproic_acid
simvastatin
isoniazid
nitrosodiethylamine
quinidine
enalapril
clofibrate
naproxen
theophylline
Log2(FC)
low
middle
high
Concentration


sulfasalazine
enalapril
cephalothin
papaverine
simvastatin
disopyramide
caffeine
naproxen
diclofenac
isoniazid
naphthyl_isothiocyanate
gentamicin
ibuprofen
theophylline
colchicine
Log2(FC)
Concentration
low
middle
high

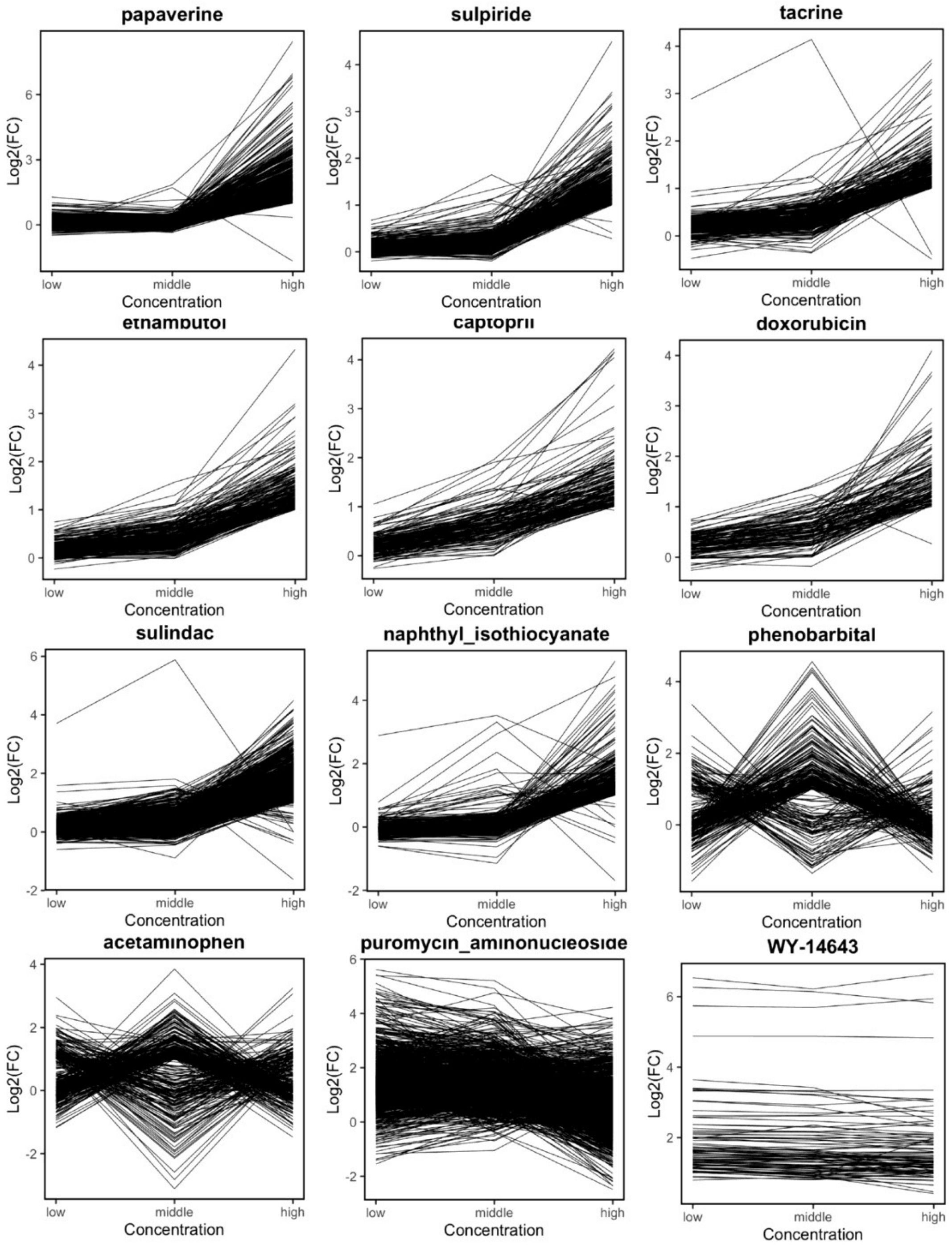


Figure 3.4 Expression levels of the most affected genes upregulated at three concentrations after 24h exposure

3.4.2 Gene expression profiles reflect common pharmacology

We clustered compounds according to an overall measure of gene expression similarity (considering all genes dysregulated at any dose or time point equally) and inspected the resulting groups for similarities in action of drug, toxicity, or drug target.

3.4.2.1 *In vivo* gene expression clusters

The dendrogram of *in vivo* gene expression data includes 98 out of 131 compounds that had signatures with at least ten and at most 200 significantly changed genes (**Figure 3.5**). Compounds are in three main clusters with multiple sub-clusters. Several sub-clusters (enclosed by boxes in **Figure 3.5**) show similar ATC classifications or drug targets and suggesting that drug action is a major determinate of gene expression. Only two sub-groups show a clear clustering according to toxicity (hepatotoxicity).

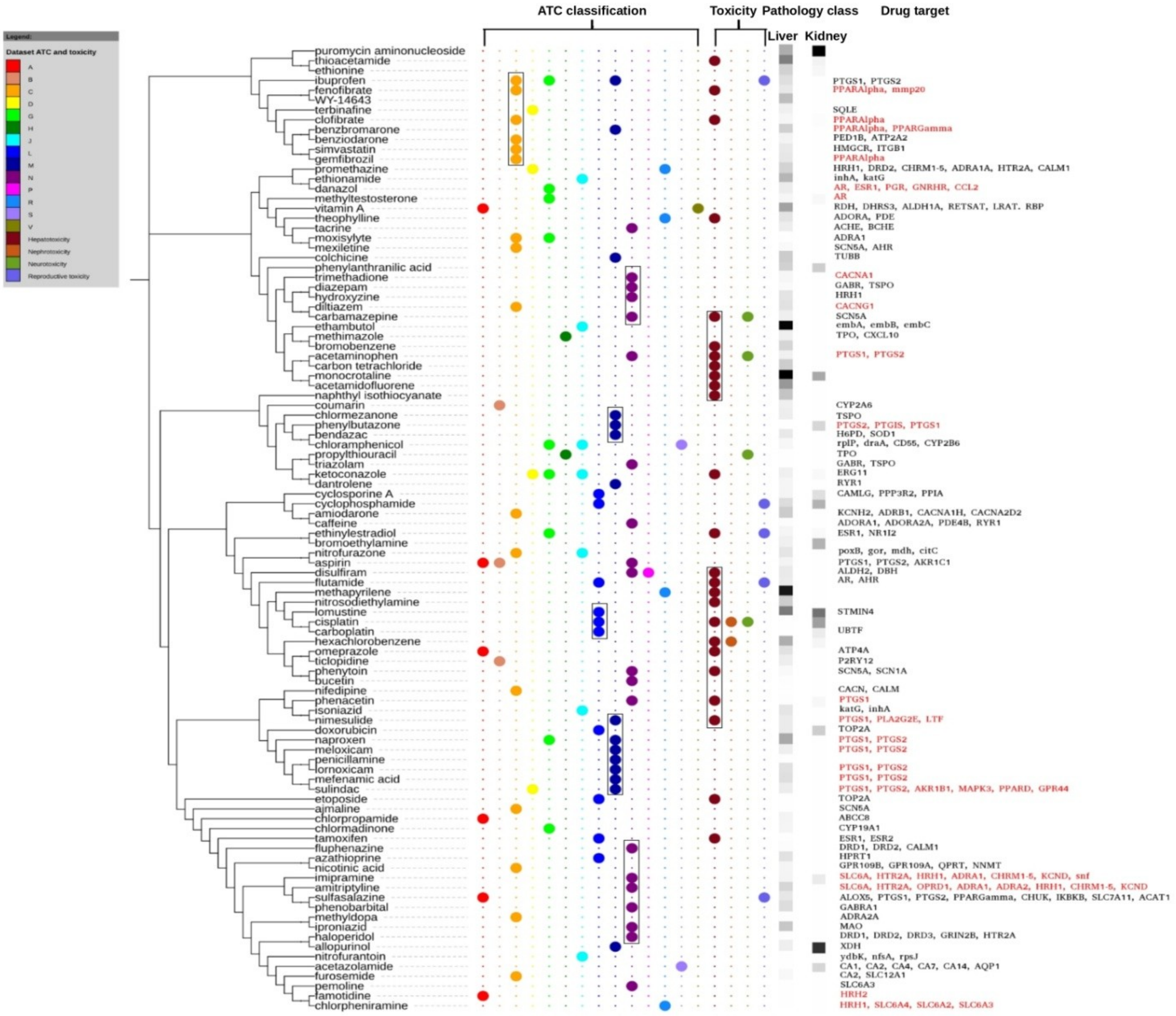
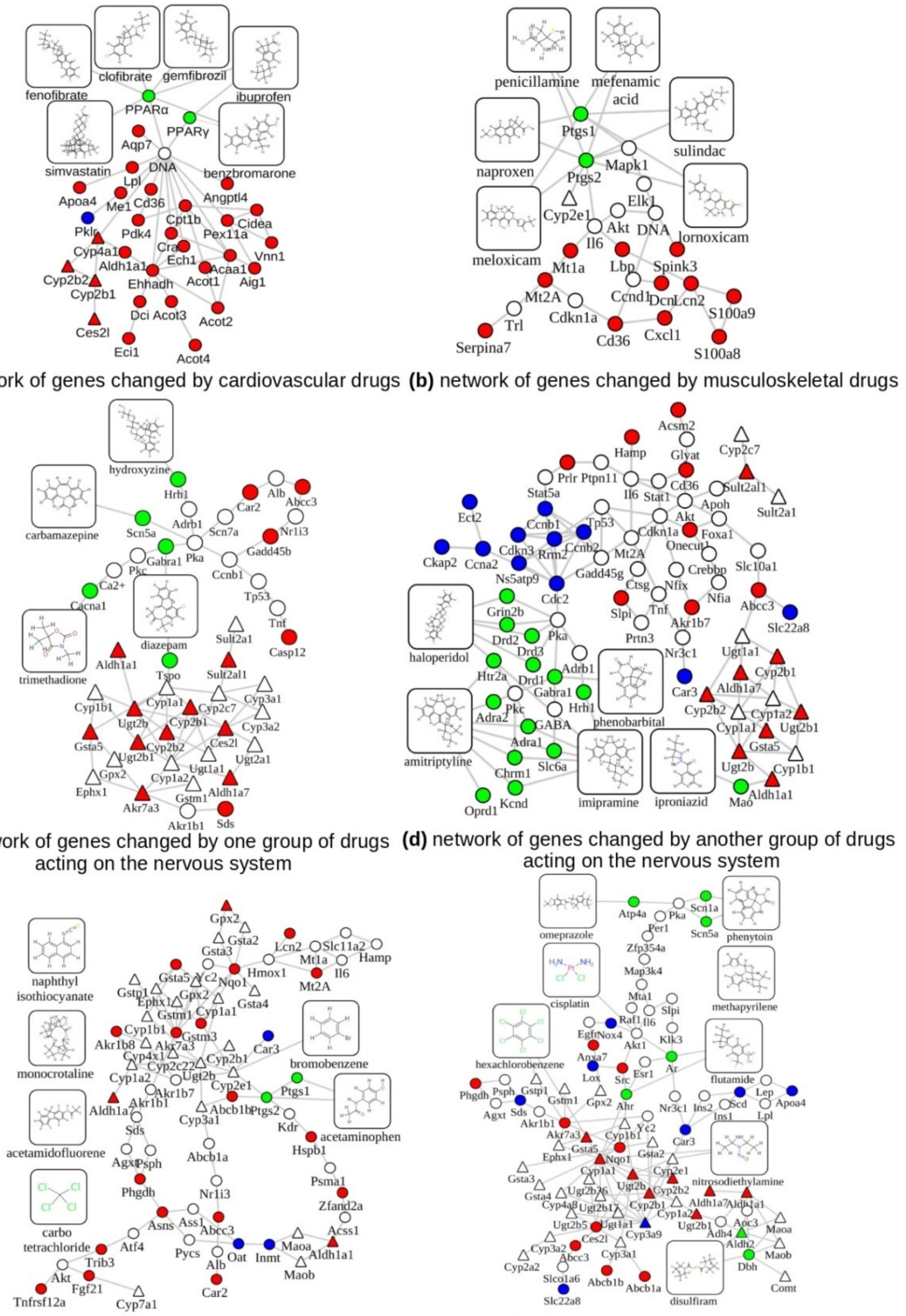


Figure 3.5 Dendrogram showing the clustering of 98 compounds based on gene regulation *in vivo* rat liver. The enriched biological features are labeled on the tree. The exact black color displays highest value of pathology classes and the colors gradually fading towards white representing the lower values.



(a) network of genes changed by cardiovascular drugs **(b)** network of genes changed by musculoskeletal drugs

(c) network of genes changed by one group of drugs acting on the nervous system

(d) network of genes changed by another group of drugs acting on the nervous system

(e) network of genes changed by one group of hepatotoxicants

(f) network of genes changed by another group of hepatotoxicants

Figure 3.6 Networks of genes changed by chemicals. Nodes are distinguished by different shapes and colors including compounds represented by 2D structures (round rectangles), drug targets (green ellipses), drug metabolizing genes (triangles), genes (ellipses), up-regulated genes (red ellipses), down-regulated genes (blue ellipses), and drug targets (green ellipses) and edges includes drug-target association, target-pathway association and gene-gene association. **(a)** network of genes changed by cardiovascular drugs (e.g.

benzbromarone, clofibrate, fenofibrate, gemfibrozil, ibuprofen and simvastatin) **(b)** network of genes changed by musculoskeletal drugs (e.g. naproxen, meloxicam, penicillamine, lornoxicam, mefenamic acid and sulindac) **(c)** network of genes changed by one group of CNS drugs (e.g. trimethadione, diazepam, carbamazepine and hydroxyzine) **(d)** network of genes changed by another group of CNS drugs (e.g. amitriptyline, imipramine, phenobarbital, haloperidol and iproniazid) **(e)** network of genes changed by one group of hepatotoxicants (e.g. acetamidofluerene, monocrotaline, carbo tetrachloride, naphthyl isothiocyanate, acetaminophen and bromobenzene) **(f)** network of genes changed by another group of hepatotoxicants (e.g. hexachlorobenzene, cisplatin, methapyrilene, nitrosodiethylamine, disulfiram, flutamide, phenytoin and omeprazole)

There is one large sub-group of predominantly cardiovascular agents (ATC code: C): ibuprofen, fenofibrate, clofibrate, benziodarone, simvastatin and gemfibrozil. These drugs target either PPARA or PPARG, and nine of the common dysregulated genes are likely altered by PPAR transcriptional activation/suppression events. Twelve of 29 of the genes changed by at least four agents in this cluster are involved in peroxisome proliferation events (*Cd36*, *Ehhadh*, *Pex11a*, *Me1*, *Acaa1*, *Angptl4*, *Crat*, *Cyp4a1*, *Ech1*, *Aqp7*, *Cpt1b* and *Lpl*), together with fatty acid metabolism, including genes for fatty acid elongation/degradation and lipid metabolism (e.g., *Acot1*, *Acot2*, *Acot3*, *Acot4*, *Dci*, *Eci1* and *Apoa4*) (**Table S3.6**, **Figure 3.6(a)**).

The sub-cluster of musculoskeletal drugs (ATC code: M) comprises six anti-inflammatory drugs (naproxen, meloxicam, penicillamine, lornoxicam, mefenamic acid and sulindac), with similar expression patterns. 14 genes were changed by at least five of those compounds (**Table S3.7**). The mechanism of anti-inflammatory drugs is mainly driven by Ptgs1 and Ptgs2, and genes centering around the pathway of anti-inflammatory were up- or down-regulated by all the six compounds, such as *Cxcl1*, *Cd36*, *Mt1a*, *Mt2A*, *Lbp* and *Spink3* (**Figure 3.6(b)**). Another sub-cluster contains three additional anti-inflammatory drugs, phenylbutazone, bendazac and chlormezanone, with eight common up-regulated genes affecting retinol metabolism, glutathione depletion and drug metabolism (*Abcc3*, *Aldh1a1*, *Aldh1a7*, *Cyp2b1*, *Cyp2b2*, *Gadd45b*, *Gsta5*, and *Rgd1562844*; **Table S3.7**).

A group of central nervous system (CNS) drugs (ATC code: N) lie in two subclasses. The first group includes trimethadione, diazepam, hydroxyzine and carbamazepine that all have been demonstrated to enhance the expression of drug-metabolizing genes, such as aldehyde dehydrogenase (*Aldh1a1* and *Aldh1a7*), cytochrome P450 (*Cyp2b1* and *Cyp2b2*), UDP glucuronosyltransferase (*Ugt2b* and *Ugt2b1*), glutathione S-transferase alpha 5

(*Gsta5*) and carboxylesterase 2 (*Ces2c* and *Ces2l*) (**Table S3.8**, **Figure 3.6(c)**). Two of those compounds, trimethadione and carbamazepine, commonly affected all 15 genes in the network, while others affected nine. Compounds in the other CNS drugs cluster containing imipramine, amitriptyline, phenobarbital, iproniazid and haloperidol) often (at least three) dysregulated *Cdc2*, *Cdkn3*, *Rrm2*, *Ccna2*, *Ccnb1*, *Ccnb2*, *Ckap2*, *Ns5atp9* and *Ect2*, which are related to the cell cycle, and up-regulated genes for neuroactive ligand-receptors, such as *Prlr*, in addition to drug-metabolizing enzymes (**Table S3.8**, **Figure 3.6(d)**). One of the similarities between the two networks (**Figure 3.6(c)** and **Figure 3.6(d)**) is that genes dysregulated by all CNS drugs interact with protein kinase A (PKA), suggesting that their therapeutic mode of action is at least, in part, linked to the cAMP/PKA signaling pathway.

There are also some compound clusters where common differentially expressed genes correlated roughly with a similarity in toxicity. For instance, bromobenzene, acetaminophen, carbon tetrachloride, monocrotaline, acetamidofluorene and naphthyl isothioxyanale are all hepatotoxic and at least four of them affected genes strongly associated with liver toxicities (**Table S3.9**). These include several genes encoding drug metabolizing enzymes and transporters (*Aldh1a1*, *Akr7a3*, *Asns*, *Gpx2*, *Aldh1a7*, *Gsta5*, *Abcb1a*, *Abcb1b* and *Abcc3*; **Figure 3.6(e)**). These principally work as protective enzymes responding to oxidant stress in phase II or III drug metabolism (Tanaka et al., 2007). Similarly, in the cluster including disulfiram, flutamide, methapyrilene, nitrosodiethylamine, cisplatin, hexachlorobenzene, omeprazole and phenytoin, both affect the genes for the first group in addition to down-regulating genes related to lipoprotein secretion (*Apoa4*, *Scd* and *Scd1*) (**Figure 3.6(f)**). The inhibition of lipoprotein secretion is known to play a role in hepatocellular necrosis and steatosis (Gao et al., 2010). Aside from these examples, the grouping of compounds according to gene-expression similarity shows little overall overlap with observed pathology outcomes (black boxes in **Figure 3.5**).

3.4.2.2 *In vitro* gene expression clusters

The dendrogram (64 compounds) for *in vitro* rat liver (**Figure 3.7**) also shows clusters more evidently related to drug action than toxicity. Some clusters resemble the *in vivo* conditions, such as that for cardiovascular drugs (simvastatin, benziodarone and clofibrate), CNS drugs (imipramine, iproniazid and amitriptyline), and alkylating agents (cisplatin and carboplatin).

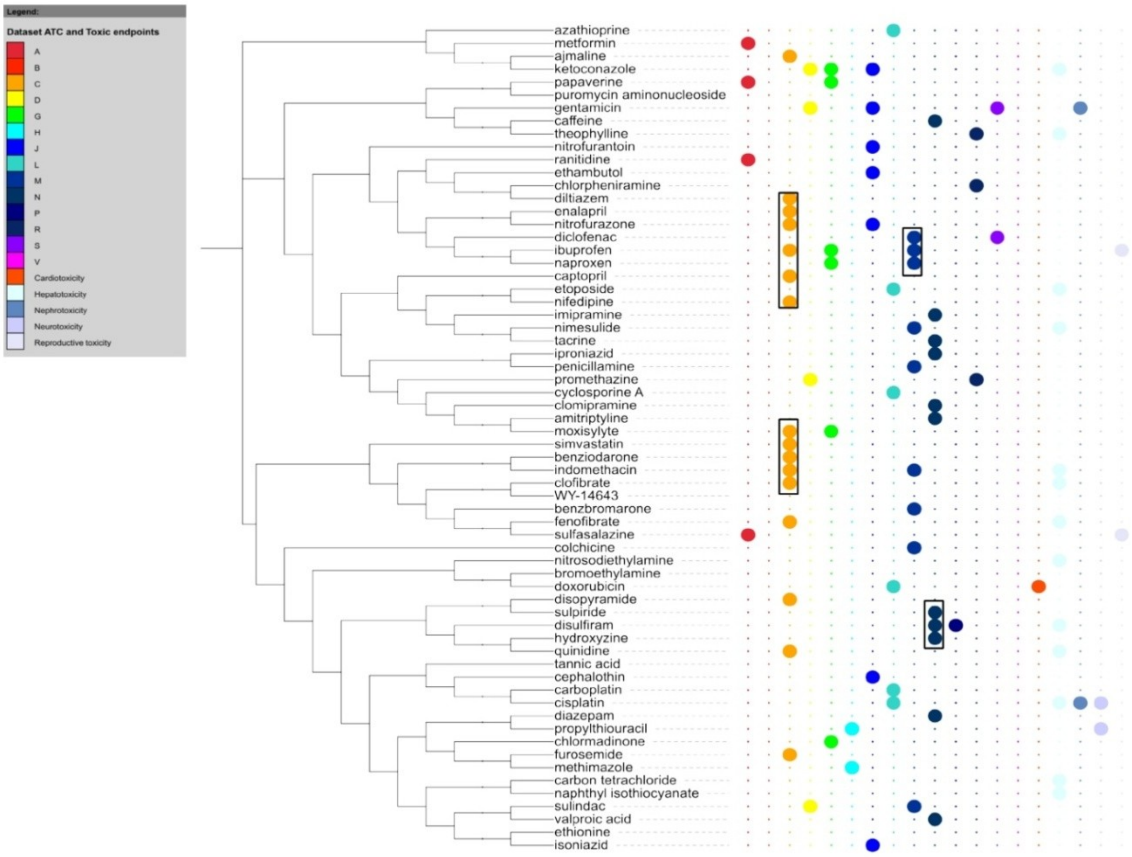


Figure 3.7 Dendrogram showing the clustering of 64 compounds based on gene regulation *in vitro* rat hepatocytes. The enriched biological features are labeled on the tree.

In *In vitro* array system only three genes related to the PPAR signaling pathway (*Acot1*, *Ehhadh*, and *Angptl4*) were also induced by cardiovascular drugs *in vivo*. *Aldh1a1*, *Acot3*, *Apoa4*, *Crat* and *Pklr* were only modified *in vivo* but not *in vitro*, while PPAR related genes *Hmgcs2*, *Cyp4a2*, *Cyp4a3* were induced *in vitro* only (**Table S3.6**). The clustered CNS drugs, imipramine, iproniazid and amitriptyline did not share any common dysregulated genes *in vivo* and *in vitro* (**Table S3.8**).

Additional clusters also reflect common drug targets. For instance, caffeine and theophylline target adenosine receptor antagonists; diclofenac, ibuprofen and naproxen target Ptgs1/Ptgs2; enalapril and captopril interact with angiotensin I converting enzyme (Ace); clomipramine and amitriptyline share multiple targets such as neurotransmitter transporter (Slc6a2 and Slc6a4) and 5-hydroxytryptamine receptor 2A (Htr2a); and clofibrate, benzbromarone and fenofibrate target PPARA.

There are also some sub-clusters reflecting common hepatotoxicity as for the *in vivo* data, such as the clusters of indomethacin and clofibrate, of disulfiram and quinidine, and of carbon tetrachloride and naphthyl isothiocyanate. The first cluster shows increased some of the same gene expressions of PPAR signaling pathway as the *in vivo* data (e.g. *Acot1*, *Acot2*, *Angptl4*, *Cyp4a2*, *Cyp4a3*, *Ehhadh* and *Pdk4*). The second cluster also shows dysregulation of *Crct1*, *Cyp1a1* and *Tcp11l2*, and the last cluster also induced genes associated with MAPK signaling pathway such as *Hspa1a*, *Hspa1b*, *Hspb1* and *Il6* (**Table S3.9**).

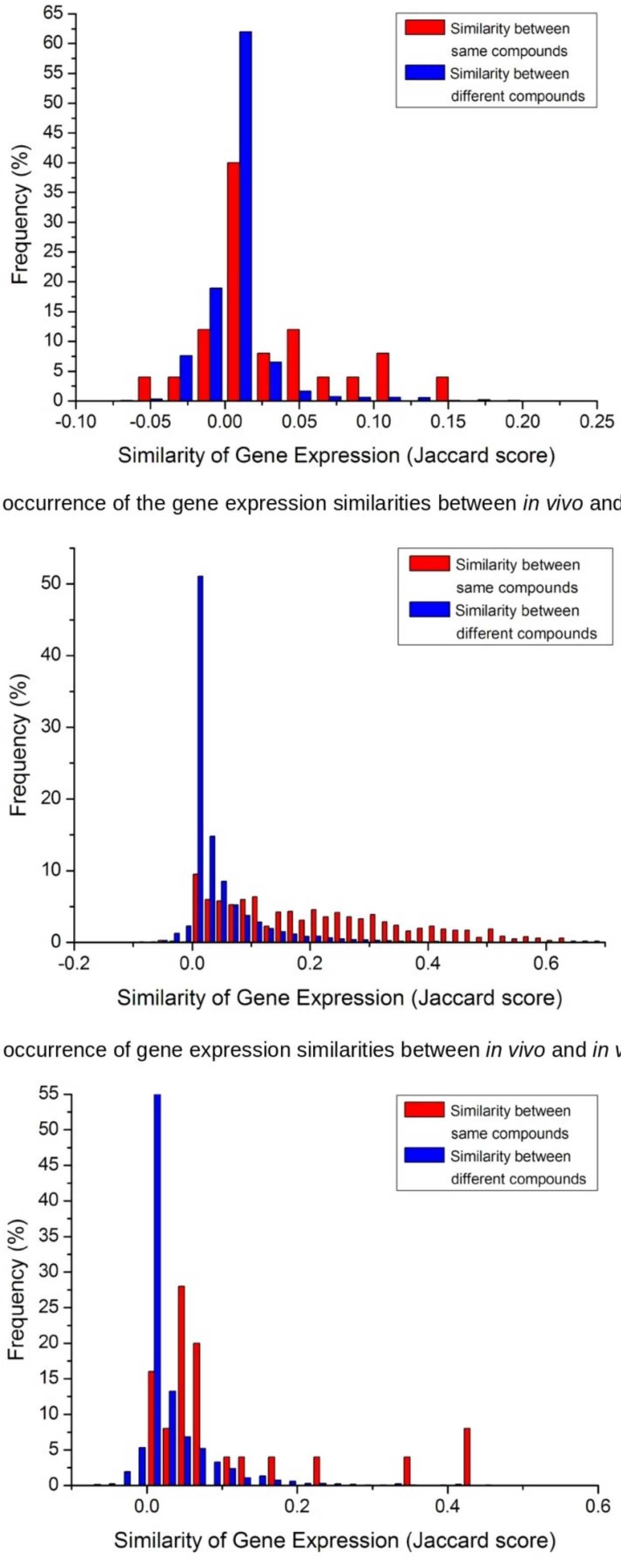
3.4.3 Comparison of *in vivo* and *in vitro* gene expression

In *in vivo* assay system, all 131 compounds changed expression of at least one gene at each dose level and time point while in *in vitro* rat hepatocytes 28 compounds changed none. In some instances, this could have a profound impact on interpretation of results. For example, considering genes modified by PPARA ligands (**Table S3.6**), seven of the fatty acid *beta* oxidation-related genes were commonly induced *in vivo* rat liver and *in vitro* rat hepatocytes by at least five agents, whereas genes responsible for cell communication (*Acta1*) and cytochrome P450 (*Cyp1a1*, *Cyp4a2* and *Cyp4a3*) and PPAR signaling pathways (*Hmgcs2*, *Mmp1*) were altered *in vitro* rat hepatocytes but not *in vivo* rat liver. There are no overlapping genes dysregulated by three drugs of the clustered CNS drugs, imipramine, iproniazid and amitriptyline between *in vitro* rat hepatocytes and *in vivo* rat liver (**Table S3.8**). With the treatments of cisplatin and carboplatin from the alkylating cluster, *Aldh1a1*, *Ccng1*, and the selenium pathway genes *HbaA2*, *Hba1*, *Hba2*, and *Lox*, *Rt1Ce5*, *Rt1Ec2* were commonly changed *in vivo* rat liver but not *in vitro* rat hepatocytes, while *Aif1*, *C1qa*, *Cd48*, *Cxcr4*, *Gja1*, *Marcks*, *Rgs4* and *Slc4a11* were exclusively changed *in vitro* rat hepatocytes (**Table S3.10**).

Expression changes *in vitro* thus might not always reflect the situation *in vivo.* To quantify this, we computed a simple measure of gene expression similarity and studied the differences between the *in vitro* rat hepatocytes and *in vivo* rat liver expression signatures (**Figure 3.8**). The most similar signatures were those for clofibrate where twelve genes show a similar behavior out of 105 genes changed *in vivo* rat liver and 29 *in vitro* rat hepatocytes*,* of which ten are clearly involved in PPAR signaling and fatty acid metabolism.

Similarities between the same compound when compared across *in vivo* rat liver and *in* *vitro* rat hepatocytes *(***Figure 3.8 (a)**) are higher than between different compounds and this is also generally true when comparing signatures (at different time points) within *in vivo* rat liver (**Figure 3.8 (b)**) and within *in vitro* rat hepatocytes (**Figure 3.8 (c)**) separately. However, there are many compounds for which similarity when comparing signatures with themselves (*in vivo* versus *in vitro*) is as low as that when comparing with other compounds.

All of these plots suggest that there is a limited general correlation between gene expression alterations induced by rat liver *in vivo* and in cultivated hepatocytes. However, certain key genes modified *in vitro* might be reliably used to predict the changes *in vivo*. For example, numerous pathways such as focal adhesion, adipocytokine/PI3K-Akt/NF-κB signaling, cell cycle related to inflammatory responses and cell growth were initiated by penicillamine stimulation at 29 days *in vivo* (**Figure 3.9(a)**). Among those, adipocytokine and PI3K-Akt signaling pathway were activated by genes of *Cd36* and *Cxcl1*, respectively. These pathways lead to inflammatory responses and accordingly the detoxification gene of metallothionein (*Mt2A*) induced and genes of *Onecut1* and *Rrm2* involved with cell cycle reduced. Interestingly, all genes associated with the inflammatory state significantly changed by penicillamine at 24 hours *in vitro* were contained in those pathways although such genes (*Ccna2*, *Ccnb1*, *Igfbp3*, *Dusp9*, and *Rrm2*) were all suppressed, which may reflect the similar mechanism but a different cell responses possibly because of dose differences, i.e., high concentration penicillamine might already induce toxicity *in vitro* (**Figure 3.9(b)**).



(a) occurrence of the gene expression similarities between *in vivo* and *in vitro*

(b) occurrence of gene expression similarities between *in vivo* and *in vivo*

(c) occurrence of gene expression similarities between *in vitro* and *in vitro*

Figure 3.8 Histogram of gene expression similarities from all-by-all compound comparison. The column highlighted in red shows the frequency which counts for gene expression values are at least 2 fold changed

induced by any two same compounds, and the blue column shows the frequency of that induced by any two different compounds. **(a)** occurrence of the gene expression similarities between *in vivo* and *in vitro* **(b)** occurrence of gene expression similarities between *in vivo* and *in vivo* **(c)** occurrence of the gene expression similarities between *in vitro* and *in vitro*

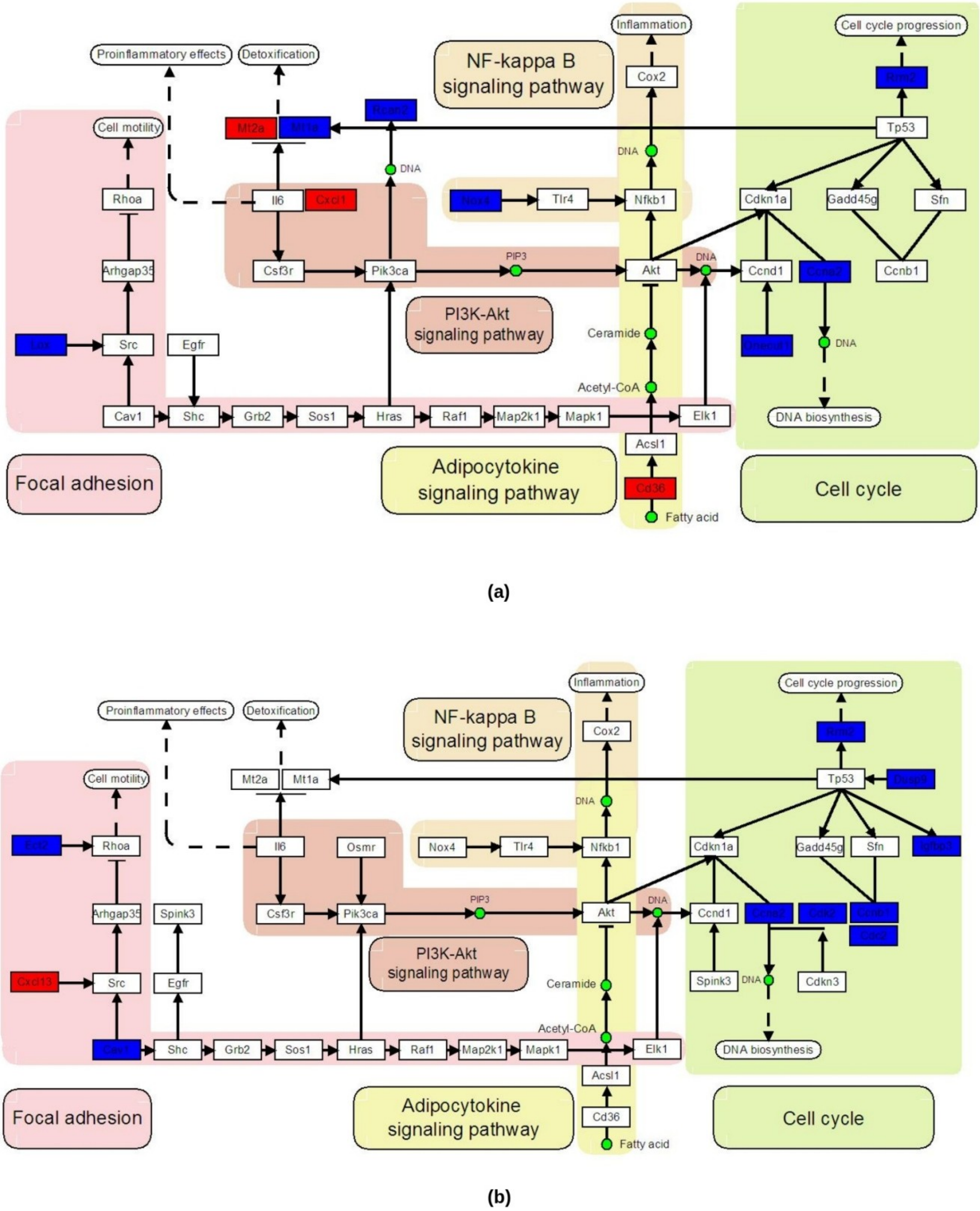


Figure 3.9 Network structure presentation of fold change. **(a)** The fold change data of rat liver at 29 days after penicillamine treatment are presented, where red and blue indicate up (fold change > 2, p < 0.01) and down (fold change < -2, p < 0.01), respectively. **(b)** The fold change data of rat hepatocytes at 24 hours after penicillamine treatment are presented, where red and blue indicate up (fold change > 2, p < 0.01) and down

(fold change < -2, p < 0.01), respectively.

3.4.4 Comparison between Rat & Human data

We studied microarray data for drug-treated human and rat hepatocytes *in vitro* with the purpose of understanding how well human cell responses to chemicals can be inferred from rat. The cluster tree of gene expression also indicates that several chemicals with similar ATC code gather into small subclasses (**Figure 3.10**). However, only 9.8% of these genes were common for human and rat hepatocytes and 11% were conserved between *in vitro* human hepatocyte and *in vivo* rat liver.

None of the genes related to the *beta* oxidation of fatty acids were changed in human hepatocytes (**Table S3.6**), which agreed with the previous observations that PPARα ligands do not cause peroxisome proliferation in humans (Harmon et al., 2011). Similarly, the changes of genes treated with compounds belonging to CNS agents, musculoskeletal drugs, antineoplastic and immunomodulating agents were not apparent in human hepatocytes (**Table S3.8**, **S3.7**, **S3.10**), but genes *ANLN*, *CDC2*, *NCAPG*, *RRM2* and *TOP2A* involved in the cell cycle were consistently down-regulated by alimentary tract drugs in human hepatocytes (**Table S3.11**), which might indicate that these chemicals have induced toxicity in human hepatocytes.

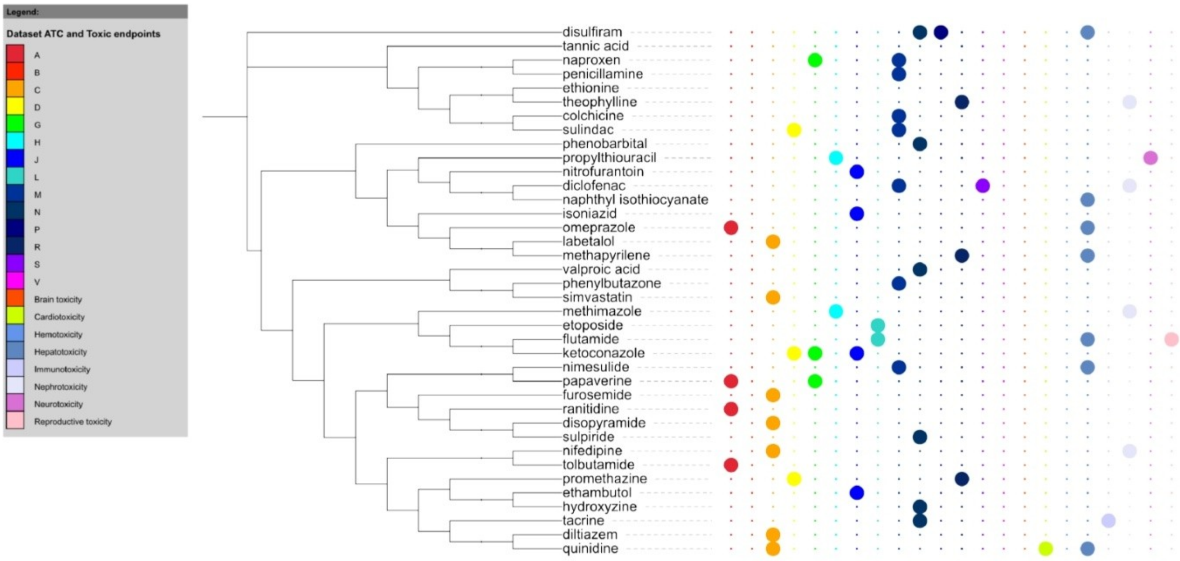


Figure 3.10 Dendrogram showing the clustering of 38 compounds based on gene regulation *in vitro* human hepatocytes. The enriched biological features are labeled on the tree.

3.4.5 Liver and kidney gene expression comparison *in vivo*

To test how tissue selection affects the gene expression patterns of chemicals, we compared gene expression data obtained from liver and kidney. Unsurprisingly, compound administration resulted in relatively few gene expression changes at any time point in the kidney compared with the liver. Increases in the expression levels of numerous genes related to immune response were more pronounced in rat kidney than liver *in vivo* for allopurinol, gentamicin, cephalothin, enalapril or triamterene. Elsewhere, the class of antineoplastic and immunomodulating agents occupies a large part in the cluster tree of gene expression (**Figure 3.11**) which universally elevated expression of renal damage genes, e.g., *Snca*, *Cdkn1a* and *Havcr1* (**Table S3.12**). These differences suggest that gene expression profiling involving multiple tissues might be helpful in identifying the target tissues of chemicals with unknown mechanisms of therapy and toxicity.

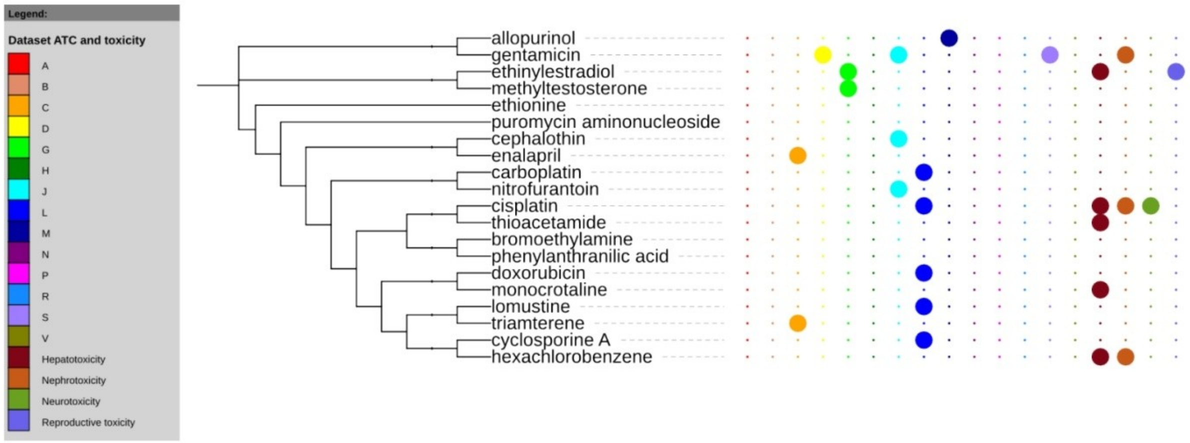


Figure 3.11 Dendrogram showing the clustering of 20 compounds based on gene regulation *in vivo* rat kidney. The enriched biological features are labeled on the tree.

3.4.6 Agreement between gene expression profiles and chemical features

3.4.6.1 Correlation of gene expression patterns and molecular descriptors

We studied the impact of a set of basic chemical descriptors of its structure and function across 98 compounds on the *in vivo* gene expression. The key assumption is that the chemical properties as encoded in the descriptors representing the lipophilicity of compounds impacts on the specific patterns of gene expression. In more specific terms, the drug metabolizing genes *Aldh1a1*, *Aldh1a7*, *Cyp2b1*, *Abcc3*, *Akr7a3* and *Ces2l* were the top six genes most often up-regulated by the 80 lipophilic agents *in vivo*, whereas the top six cell proliferation genes (*Cdk1*, *Cdkn3*, *Rrm2*, *Onecut1*, *Stac3* and *Cish*) were mostly up-regulated in expression by 15 hydrophilic compounds and are all involved in the anti-inflammatory pathway (**Table S3.13**). This agrees broadly with the notion that lipophilic and hydrophilic drugs differ in their clinical capabilities. For example, lipophilic xenobiotic compounds are typically converted/activated by drug metabolizing enzymes in phase I (Hodgson and Goldstein, 2001; Iyanagi, 2007) and then conjugated with endogenous hydrophilic molecules (sugars, glutathione) resulting in hydrophilic forms by Phase II enzymes (Iyanagi, 2007), while the hydrophilic xenobiotic containing functional groups can directly enter Phase II (Eapen et al., 2007).

3.4.6.2 Correlation of gene expression similarity and molecular similarity

We computed the pairwise correlation between Jaccard coefficient of the gene expression fold change *in vivo* or *in vitro* and their corresponding molecular similarity, separately. We then fit linear model with the similarity of the gene expression as the independent variable and the similarity of the chemical structures as the dependent variable. Interestingly, the molecular similarity (Tanimoto, which is also a Jaccard index) of the chemicals is not correlated well with similarity in gene expression (**Figure 3.12**), though most points correspond to low similarities of both gene expression and molecular structure. There are individual instances showing higher gene expression similarity with lower molecular similarity and higher molecular similarity with lower gene expression similarity (**Table S14**). Virtually all instances of highly similar gene expression with no molecular similarity correspond to PPARα ligands (with increased lipid metabolizing genes). It is well known that PPARα ligands can lack any apparent molecular similarity (Harmon et al., 2011), and only the fibrates have clearly similar molecular features (**Figure 3.13**).

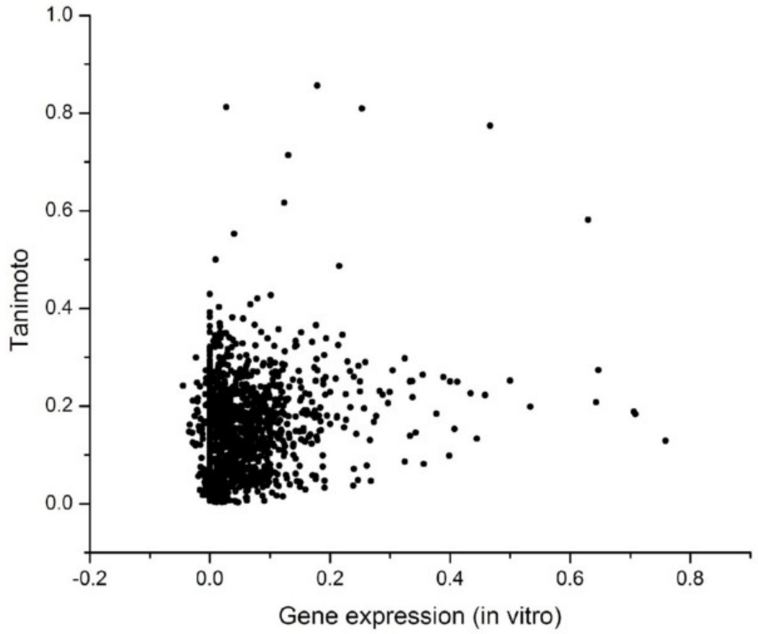


Figure 3.12 Correlation of similarity between gene expression and molecular structure.

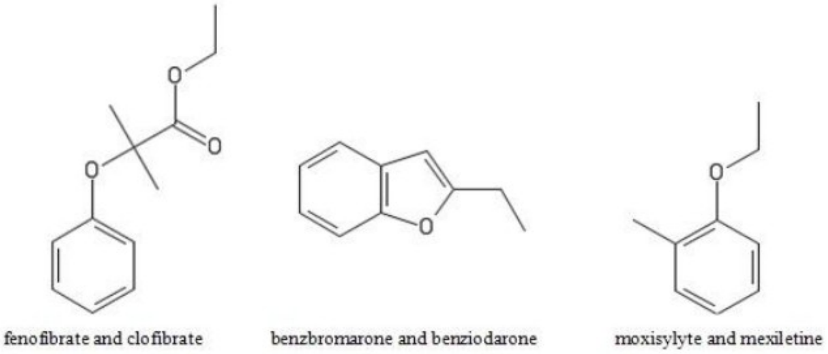

O
O
O
O
fenofibrate and clofibrate
benzbromarone and benziodarone
moxisylyte and mexiletine

Figure 3.13 Some common fragments found in compounds with very similar gene expression patterns in rat liver.

3.4.6.7 Correlation of gene expression and chemical fragments

The fragment matrix (98 compounds x 1332 fragment identifiers (> 8 non-hydrogen atoms)) shows a projection of fragments into a reduced chemical space. We used PCA to reduce these 1332 fragments to 135 loadings on the first two principal components. The plot of the first two components (**Figure 3.14**) shows a separation of cardiovascular, musculoskeletal, and CNS drugs and a cluster of antineoplastic and immunomodulating agents.

The most prevalent fragments in each class are listed in **Figure 3.15**. Acetanilide substructure containing compounds (first cluster), including diazepam, diltiazem, acetaminophen, acetamidoflurorene and phenylbutazone, commonly induced drug transporters (*Abcb1a*, *Abcc3*) and drug metabolizing genes (*Akr7a3*, *Aldh1a1*, *Cyp2b1*, *Sult2al1* and *Ugt2b1*). Another two acetanilide containing compounds, bucetin and phenacetin, induced (in addition to these drug-metabolising genes) L-serine biosynthesis genes such as *Phgdh* (phosphoglycerate dehydrogenase) and *Psat1* (phosphoserine aminotransferase), which were reported to be related to cell proliferation (Sun et al., 2015). Compounds containing the 2-(2-methylphenoxy) moiety (ethanamine, mexiletine and moxisylyte), all up-regulated genes in the estrogen and PI3K-Akt signaling pathways. Chlormezanone, ketoconazole, chlorpheniramine, haloperidol and furosemide contain a scaffold of 4-chlorotoluene (second cluster) and commonly dysregulated cell proliferation related genes, including *Cish* and *Gadd45b*. Furan-2-carbaldehyde-hydrazone containing compounds (dantrolene, nitrofurazone and nitrofurantoin) (third cluster) all increased drug transport and metabolism genes *Abcc3* and *Akr7a3* and decreased the cell proliferation gene *Rrm2*.

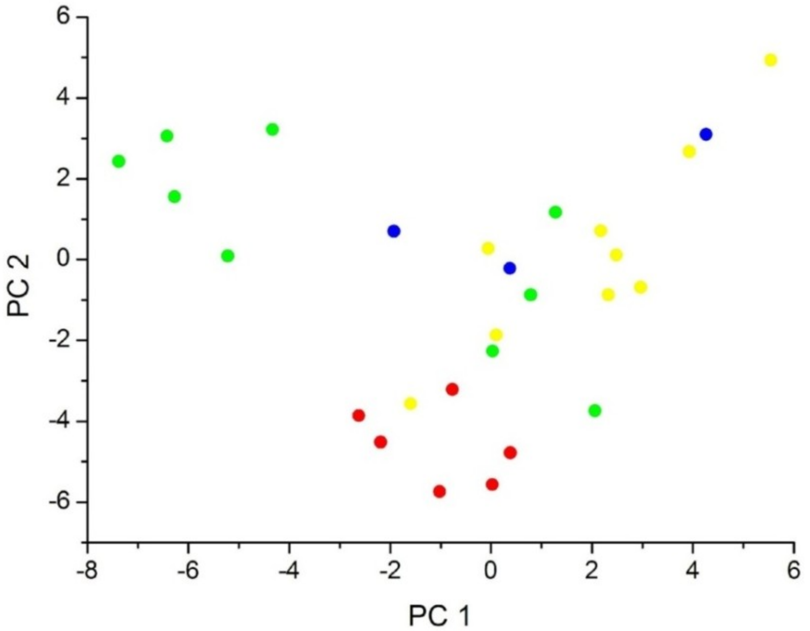
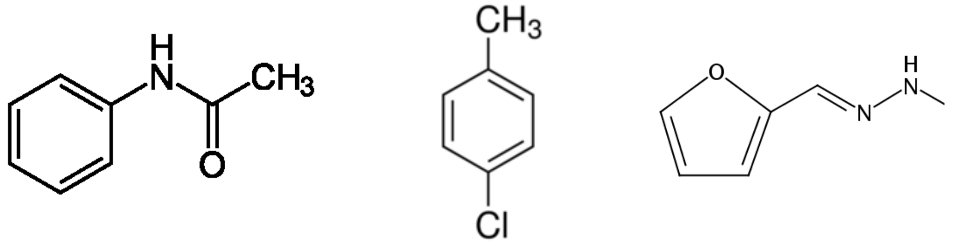


Figure 3.14 Fragment space of chemicals defined by the first two factors from an analysis of 1332 substructures. Colors highlight different clusters of drugs (red - first cluster – cardiovascular drugs; yellow - second cluster – musculoskeletal drugs; green – third cluster – CNS drugs; blue – fourth cluster - antineoplastic and immunomodulating drugs)



acetanilide **4-chlorotoluene** **furan-2-carbaldehyde**

Figure 3.15 Most prevalent fragments found in three clusters of chemicals.

3.4.7 Predicting pathology signatures based on chemical fragments and gene expression data

3.4.7.1 Correlation analysis between gene expression and pathology signatures

We used TG-GATES histopathology data of liver and kidney to build matrices containing 131 compounds and 31 pathological terms for *in vivo* rat liver, 24 for *in vivo* kidney, and 31 for rat hepatocytes (mapped from *in vivo* data). Each compound profile has elements of 1 or 0 denoting whether or not it is associated with each pathological term. To encode the gene expression profiles of 131 compounds, genes dysregulated by at least four compounds at any time points and any dose levels were considered (FC > 2, or FC < -2, and p < 0.01). Each compound was thus represented by 853, 396 and 966 binary values for liver, kidney and hepatocytes.

Gene expression and histopathological records were randomly split into a training set (80%) and a test set (20%). SCCA was trained by 5-fold cross-validation using R (Pauwels et al, 2011). The performance was evaluated by ROC curves (**Figure 3.16** all predictions merged into one curve). SCCA was computed using parameters of c1 = c2 = 0.04 (penalty) and m = 20 (number of components). The resulting AUC scores are 0.7883, 0.731, and 0.7545. The curves suggest that there is capable predictive power of using these data.

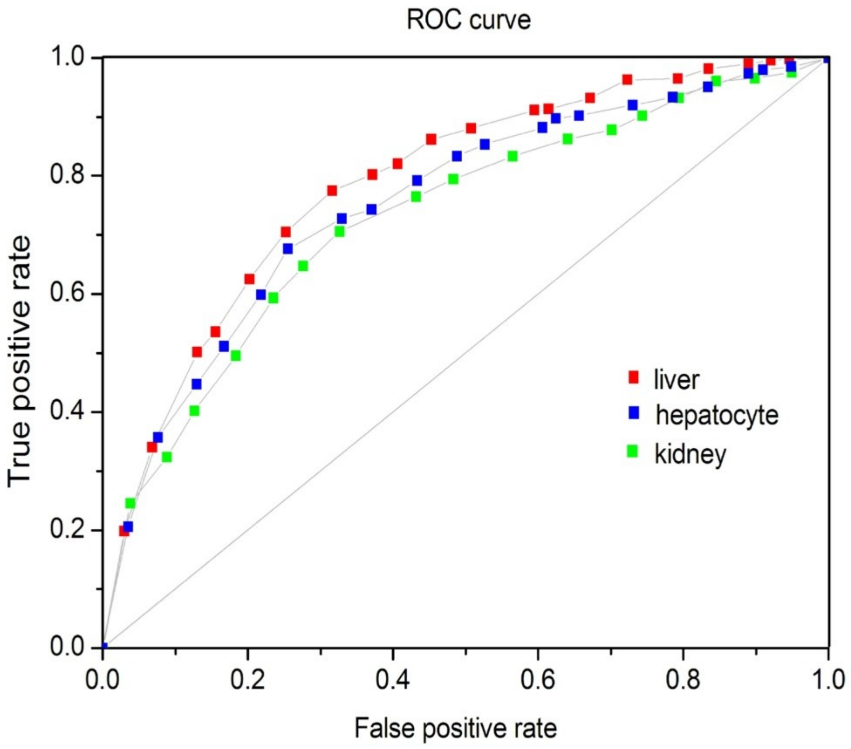


Figure 3.16 ROC curves for the 5-fold cross-validation comparing the performance of SCCA on *in vivo* rat liver, kidney and *in vitro* rat hepatocytes

We examined the weight vectors for chemical fragments and pathological outcomes in SCCA to extract the relationship between two datasets. **Table S3.15** shows the weight vectors of canonical components in SCCA and most of the elements in the weight vectors are zero, implying that SCCA can select a small number of gene-sets or pathologies as informative. A few sets of genes and pathological records in the canonical component with correlation coefficient larger than 0.6 were extracted using SCCA. Each component consists of an association of a set of genes and one pathological term. For each component, two lists of compounds are provided, one containing compounds with a high score for the associated genes, and the other containing compounds with a high score for the associated pathological observations. **Table S3.16**, **S3.17**, **S3.18** show the selected correlations between genes and pathological observations for liver, kidney and hepatocyte experiments, separately.

Several of these components show plausible biological mechanisms. For instance, component 2 corresponds to "inflammation" and includes a set of up-regulated genes, such as *Il7*, *Ccl21*, *Stat3*, *Dsc2* and *Slc13a5* that are involved in inflammation pathways in the liver dataset. The top ten compounds scored by this cluster belong to the same ATC inflammation category (**Table S3.16**). This finding agrees with our earlier conclusion that gene expression patterns tend to reflect the MoA of drugs rather than toxicity *per se*.

3.4.7.2 Correlation analysis between chemical fragments and pathological signatures for *in vivo* liver

31 histopathological terms for 52 compounds were extracted from TG-GATEs and 25 chemical fragment contained by those compounds were defined by the CCQ fragmentation method (ChemAxon, https://www.chemaxon.com), which leaves functional groups of molecules intact so that they are better reflections of potential mechanism. Each compound was represented by a 31 dimensional binary profile and a 25 dimensional binary profile whose elements encode for the presence or absence of each of the pathological terms or chemical fragments. Data were randomly split into a training set (80%) and a test set (20%). SCCA was trained by 5-fold cross-validation. The resulting AUC score is 0.6108 using parameters of c1 = c2 = 0.04 (penalty) and m = 20 (number of components).

A set of fragments and pathological records in the canonical components with a correlation coefficient larger than 0.7 were extracted (**Table S3.15**). For instance, component 1, with highest canonical correlation (0.771), associates the substructure 2-aminopyridine to a list of pathological indications of hepatotoxicity such as alteration, anisonucleosis, hemorrhage and vacuolization. The 2-aminopyridine moiety, defined as structural alert of hepatotoxicity in some antibacterial drugs due to its ability to form reactive metabolite (Stepan et al., 2011). This cluster shows alterations in genes related to anisonucleosis, including *Id1* (inhibitor of DNA binding 1), *Tp53inp1* (tumor protein p53 inducible nuclear protein 1) and *Unc5b* (unc-5 homolog B), which are all involved in apoptosis and cell death.

3.5 Conclusion and discussion

The first major finding in this chapter is that much of toxicogenomics results reflect the underlying mechanism of drug action, and indeed overall similarities in expression patterns seem to be more related to drug mechanism than underlying similarities in toxicology. This is somewhat surprising as the target tissues (liver, kidney) are not traditionally considered to be targeted by many of the drugs considered (e.g. acetaminophen, etc.). For some compounds, we also find only a weak effect on gene expression, which could reflect dosing problems as was reported previously There are, of course, exceptions to this in our own analyses (**Figure 3.1**), and indeed in the general, well established utility of gene expression analyses to uncover new toxicity biomarkers. For instance, the gene kidney injury molecular 1 (*Kim1* or *HAVCR1* in humans) is highly evident in the TG-GATES (**Table S3.12**) and other expression datasets as an indicator of kidney toxicity (Ichimura et al., 1998; Kondo et al., 2009), and many other known and potentially novel biomarker candidates are present within these and other expression data. Nevertheless, our overall findings have some important implications for methods to predict toxicity, suggesting a need to account for (or remove) mechanistic contributions to expression profiles before deriving multi-gene predictors.

The second major observation is that there is limited overlap between *in vivo* and *in vitro* studies when considered across this large set of compounds. The notion that cell lines are poor mimics of mammalian tissues is certainly not new, but our results provide a systematically derived degree of similarity for guidance. Additionally cell lines, such as 3D cell cultures that are believed to be a better mimic of human organs that 2D cultures (Fey and Wrzesinski, 2012) will likely improve upon the general picture we have seen here. The overlap among different assay systems (rat *in vivo*, rat *in vitro* and human *in vitro*) is also endpoint dependent. For example, we found that PPARα ligands such as ibuprofen, fenofibrate, and clofibrate consistently influenced lipid metabolism in all three systems. When Liu et al. limited their analysis to specific hepatotoxic endpoints, the *in vitro* to *in vivo* extrapolation potential was improved (Liu et al., 2017).

Overall, our results suggest that large scale evaluation of gene expression patterns is a useful tool for predicting mechanisms of chemotherapeutic agents and toxicity of unknown chemicals. It is apparent that each compound produces its own, unique expression profile and that similarities in profiles between compounds can indicate similarities in therapeutic and toxic mechanism. Almost certainly, a systematic analysis combined with biological and chemical features is essential to make accurate assessments of drug mechanism and toxicity based on similarity of gene expression profiles. Several studies have provided considerable sets of expression signatures to help decipher drug mechanism, toxicity or other biological features (e.g Gray, et al., 1998; Velculescu et al., 1999). The more challenging task is to relate chemical features with systemic responses. It is not always robust that chemicals with very similar structure have similar toxicogenomics profiles. However, some compounds containing the same specific substructures show similar outcomes. Therefore, combining chemical structure and toxicogenomics may improve the assessment (Low et al., 2011). Our results confirm the value of compound-centered view of phenotypes for identifying the effects of chemicals on phenotypic outcomes as suggested previously (Duran et al., 2014) and will accordingly help efforts to relate chemical structure to biological response.

CHAPTER IV: Common, functional mutations in human subpopulations potentially causing individual susceptibility to toxicities or other pathologies

4.1 Abstract

Genetic variants can determine both inter-individual and inter-ethnic differences in many physiological processes, including drug efficacy or toxicity, sensory perception, and dietary responses. Many differences are attributable to single nucleotide polymorphisms (SNPs) in key drug response, metabolism, and sensory receptor genes. In this chapter, we first study SNPs in all genes in 2 504 humans uncovered by the 1000 Genome Project to check what kinds of genes are frequently the subject of variations in healthy people. Variants are enriched in processing like the immune response, sensory perception and drug metabolism. We then explore how variants might affect these processes, particularly drug metabolism, where we find many common variants likely to alter how individuals react to xenobiotics. The results could be potentially used in personalized medicine to define different doses for discrete patient populations.

4.2 Introduction

Recent years have seen the birth of several additive/metabolite controversies. For example, the flavor enhancer monosodium glutamate (MSG) is thought to cause what was once called “Chinese restaurant syndrome” (Kwok, 1968). Tryptophan, which is high-content in turkeys and eggs, is believed to give rise to tiredness, which is an oft cited reason for fatigue after Thanksgiving dinner in the US (Hartmann and Spinweber, 1979). However, systematic studies have found nothing significant. One hypothesis is that there are small subsets of people that experience such chemical-induced syndromes, but that these fractions are not easily detected in trials involving small groups of randomly selected healthy individuals.

With the increasing availability of “-omics” technologies on the back of the human genome project, more and more studies now focus on the study of genetic variation among individuals or human populations. The most common genetic variation studied in these contexts are single nucleotide polymorphisms (SNPs), defined as mutations changing single nucleotides in at least 1% of the human population. They have a key role in genetic changes associated with diseases and individual responses to certain chemicals, such as drugs and toxins (Roses, 2000).

Recent studies of polymorphisms have demonstrated the impact of genetic variation on a many common traits. For example, there is some evidence that variations in taste or odor-perception are associated with polymorphisms in taste or olfactory receptors (Pronin et al., 2007; Jiang and Matsunami, 2015). For instance, genetic studies have identified that variations (e.g., A5T, R247H, R757C) in TAS1R3 are related to the capability of tasting MSG (umami) (Chen et al., 2009). Another example is T2R43, where the W35 allele enables individuals to taste the bitterness of the plant toxins, aloin and saccharin (Pronin et al., 2007). The first clear link between odor perception and genotype was described in olfactory receptor OR7D4, where variants cause different responses and/or preferences for food (e.g. pork) containing androstenone (Lunde et al., 2012).

Single nucleotide variations in genes encoding drug metabolizing enzymes (DMEs), drug targets, drug transporters and human-leukocyte antigen are considered to be responsible for inter-individual differences in drug response (e.g., efficacy, toxic effects) (McLeod and Evans, 2001). Many studies focus on the CYP450 gene family, where several variants are known to cause primary differences in drug-response (Puga et al., 1997). For example, for CYP2D6, there are four major phenotypes of metabolism: extensive metabolizers with two functional alleles, poor metabolizers with two non-functional alleles, intermediate metabolizers with one functional and one non-functional allele, ultra-sensitive metabolizers with one or more alleles improving enzyme activity above extensive metabolizers (Puga et al., 1997). Other variants within genes and their relevant substrates have also been studied to evaluate drug responses such as: L128R and Y139F in the warfarin resistance enzyme VKORC1 (Ma and Lu, 2011), and mutations in KRAS related to resistance to the anti-EGFR monoclonal antibodies cetuximab and panitumumab (Fakih and Wong, 2010).

Toxic effects can also result from genetic variants affecting drug pharmacokinetic/pharmacodynamic processes. For example, variations in the ryanodine receptor gene (e.g. protein change p.C614R) are found to account for susceptibility to malignant hyperthermia after anesthetic treatment in more than 50% of cases (Gillard et al., 1991). The g.C3435T variant in the drug transporter gene *ABCB1* produces an adverse reaction to anticancer agents and cardiac glycosides (Sheng et al., 2012). Analysis of variations in *GSTT1* and *GSTM1* shows that Alzheimer's disease patients with deficiencies of both genes are more likely to experience hepatotoxicity when treated with tacrine (Roy et al., 2001). Low-dose methotrexate toxicity is often found in patients with the g.C677T mutation in the drug target methylenetetrahydrofolate reductase (Spyridopoulou et al., 2012).

These examples indicate the impact of a few common variants on a common trait. However, the entire genetic landscape of inter-individual differences of response to chemicals still remains to be systemically elucidated. High-throughput sequencing has contributed to the systematic unraveling of thousands of variants potentially disturbing biological systems or causing diseases (Kilpivaara and Aaltonen, 2013). However, understanding the genetic contribution also requires knowledge of the abundance and distribution of mutations in the general population. The 1000 Genomes (1000G) Project Phase 3 (May 2013) has explored the spectrum of common human genetic variation by sequencing 2504 individuals from various populations (The 1000 Genomes Project Consortium, 2015). An earlier study from GlaxoSmithKline provided a comprehensive description of rare variants (MAF ≤ 0.5%) in human population by applying sequencing to 202 drug target genes from 14002 individuals (Nelson et al., 2012). Meanwhile, several high-throughput computational approaches to determine the influence of mutations on protein-ligand interactions have been established (Pires et al., 2014; Reva et al., 2011). Recently, we presented Mechismo (Mechanistic Interpretation of Structural Modifications) to predict whether mutations are likely to enhance or diminish the interactions of protein-protein, protein-nucleic acid and a set of protein-ion by calculating the frequency of a given residue in protein-related ligand class, which can provide the mechanistic basis of how mutations lead to specific functional sequences (Betts et al., 2015). A newer study proposes mCSM-lig, a structure-guided strategy for quantifying the influences of mutations on binding of small molecules with proteins, which completed the assessment of the effects of mutations on protein-small molecule complexes (Pires et al., 2015; Pires et at., 2016).

In this chapter, we extracted information on genetic variants in the healthy individuals from 1000 Genomes Project. We sought variations that would be expected to differentiate humans by their ability to process chemical molecules by linking the common variants (i.e. present in more than 1% population) to small molecules via chemical-protein interactions in the Mechismo system. This allowed us to predict/identify how small molecules bind to proteins and whether any common SNPs at the binding sites might impact these interactions. Some of the changes were predicted by mCSM-lig. We also linked several variants to potential individual susceptibility to toxicity for some withdrawn drugs by examining chemical-protein complexes in Mechismo. This study will have applications in the understanding of drug response across subpopulations and thus implications for personalized or precision medicine.

4.3 Methods

4.3.1 Processing Data from 1000G

1000 Genomes Project Phase 3 (May 2013 sequence freeze) data were mapped to Ensembl (Aken et al., 2016) proteins using Variant Effect Predictor (VEP) (McLaren et al., 2016) and version 79 of the human genome version GRCh37. This includes 2 504 samples from 26 human sub-populations in five major population groups: Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). Each of these populations has an associated minor-allele frequency (MAF) and there is a global value (GMAF) that spans the entire human population. The dataset also provides, for each individual, information about the zygosity of the variant (ie. whether it is absent/wild-type, heterozygous or homozygous). VEP provides predictions of the possible severity of a mutation according to SIFT version 5.2.2 (Sim et al., 2012) and PolyPhen version 2.2.2 (Adzhubei et al., 2010).

4.3.2 Deriving protein altering variants information

The source of protein altering variants in human proteins imported from the Ensembl Variation database (Chen et al., 2010) was downloaded from UniProt (UniProt Consortium, 2015). It included in total 1 271 699 unique variants. The description of GMAF for a variant was mapped from 1000G project. We defined a dataset of common missense (amino acid altering) SNPs as those with GMAF ≥ 1%. The frequency distribution of missense variants in a protein was defined as the total number of missense substitutions found within it divided by the protein length.

4.3.3 Gene enrichment analysis

For gene enrichment analysis we used GetGo (http://getgo.russelllab.org), which performs fast enrichment analysis of human gene sets considering pre-computed gene lists of different classes (Gene Ontology (Ashburner et al, 2000), Pathways, UniProt diseases, Complexes, etc). Enrichment p-values are assigned to each list using a Fisher's exact test and multiple testing corrections.

4.3.4 Relating chemical molecules to genes

We ran all the missense variants through Mechismo (mechismo.russellab.org), and chemicals at the variation sites were showed or predicted by Mechismo. Mechismo contains roughly one million connections between human proteins and connected with chemical identifiers in the protein databank (PDB) (Gutmanas et al., 2014) and the confidence of each match based on the sequence identity between the UniProt sequence and the protein of known structure. It also provided approximately 50000 protein-protein, protein-nucleic acid, protein-small-molecule interactions with known 3D structures and several million interactions identified by other methods. Since we previously benchmarked the prediction accuracy as a function of protein sequence similarity, we could use relatively low sequence similarities (Betts et al., 2015).

4.3.5 Predicting the effects of missense variations

We used mCSM-Lig to predict the impact of the missense substitutions upon the binding affinity of protein for the substrates. This method uses graph-based signature to represent the wild-type and chemical environment of a residue to predict the change upon variation in Gibb's free energy of binding or stability (Pires et at., 2016).

4.4 Results

4.4.1 Gene function enrichment

Table 4.1 shows the most enriched Gene Ontology terms and pathways for the set of 1 512 genes that have (Number of common_SNPs/length) > 0.01. There are four main groups of genes, which can also be visualized as a network (**Figure 4.1**).

Table 4.1 The most enriched Gene Ontology terms and pathways

Term/Pathway ID	Term/Pathway	Obs	Exp	*P*-value	Genes
Gene Ontology					
GO:0007608	sensory perception of smell	28	10.6	0.01016	*GJB4 GNAS OBP2A OBP2B OMP OR2D2 OR2S2 OR51B5 OR51B6 OR51I1 OR51J1 OR51M1 OR5B3 OR5D14 OR5H1 OR5H14 OR5H15 OR5H6 OR5I1 OR5K3 OR5R1 OR5W2 OR6B2 OR8D4 OR8H2 OR8H3 OR8U1 UGT2A1*
GO:0001580	detection of chemical stimulus involved in sensory perception of bitter taste	13	2.7	0.01693	*CA6 CST1 CST2 CST4 RTP4 TAS2R19 TAS2R31 TAS2R38 TAS2R4 TAS2R42 TAS2R43 TAS2R5 TAS2R7*
GO:0045095	keratin filament	27	7.2	$9.607\text{x}10^{-06}$	*KRT74 KRTAP1-1 KRTAP10-1 KRTAP10-10 KRTAP10-11 KRTAP10-3 KRTAP10-5 KRTAP10-6 KRTAP10-7 KRTAP10-9 KRTAP12-1 KRTAP12-2 KRTAP12-3 KRTAP2-1 KRTAP2-2 KRTAP2-3 KRTAP3-2 KRTAP4-1 KRTAP4-11 KRTAP4-4 KRTAP4-5 KRTAP4-7 KRTAP4-8 KRTAP4-9 KRTAP5-5 KRTAP5-9 KRTAP9-6*
GO:0006955	immune response	42	20.0	0.00148	*BPI C1QC C8A CCL11 CCL14 CCL17 CCL25 CCL4L1 CD1E CD276 CHIA CTSG FCAR FCGR1B FCGR3B GBP6 GZMA GZMB HLA-A HLA-B HLA-C HLA-DPA1 HLA-DQA1 HLA-DQA2 HLA-DRB1 IFI44L IFITM2 IGHA1 IL1RL1 IL24 IL36A IL37 IL4R IL7R KIR2DL1 KIR2DL3 KIR3DL1 LILRB2 LY75 MS4A2 PRG3 SBSPON SECTM1 TINAG TLR1 TLR10 TLR6 TNFRSF14 TRGV3 TRIM22 XCL1 XCL2*
GO:0030574	collagen catabolic process	20	5.5	0.00434	*COL11A1 COL11A2 COL17A1 COL18A1 COL19A1 COL23A1 COL2A1 COL3A1 COL4A1 COL4A2 COL4A3 COL4A4 COL4A5 COL4A6 COL5A3 COL8A2 COL9A1 COL9A2 COL9A3 MMP8*
REACTOME Pathway					
R-HSA-	antimicrobial peptides	29	6.7	$5.470\text{x}10^{-07}$	*ART1 BPI BPIFA2 BPIFB1 BPIFB4*

6803157:2					*BPIFB6 CHGA CTSG DEFA4 DEFA6 DEFB108B DEFB116 DEFB119 DEFB124 DEFB126 DEFB127 DEFB128 DEFB129 GNLY HTN3 PGLYRP2 PGLYRP3 PGLYRP4 PI3 REG3G RNASE7 S100A7A SEMG1 TLR1*
R-HSA-211859:1	biological oxidations	32	14	0.00961	*ACSM2A ACSM5 AKR7A2 AKR7A3 AKR7L CBR3 CES1 CYP1B1 CYP2A7 CYP2F1 CYP2W1 CYP3A43 CYP4A22 CYP4B1 CYP4F12 FMO2 GLYATL3 GSTA2 GSTM1 GSTM4 GSTM5 GSTZ1 MTRR NAT2 NQO2 SULT1A2 TBXAS1 UGT1A5 UGT2A1 UGT2B10 UGT2B11 UGT2B28*
R-HSA-420499:3	Class C/3 (Metabotropic glutamate/pheromone receptors)	12	2.6	0.01608	*TAS1R1 TAS1R2 TAS2R19 TAS2R20 TAS2R31 TAS2R38 TAS2R4 TAS2R42 TAS2R43 TAS2R5 TAS2R7 TAS2R9*

As might be expected, many genes are enriched in mutations in healthy people. Olfactory and taste receptors, drug metabolism, immune response, skin, hair and nails constituent (collagen and keratinization) are the functions enriched in the gene set that has more SNPs.

The most visible phenotypic variations in human - skin, hair and nails are remarkably variable. Pigmentation is a polygenic cause of this observation (Wilde et al., 2014). The high diversity might be a result of selection pressures associated with the adaption to an agriculturalist diet, and influence of UV radiation (UVR) (Wilde et al., 2014).

Odorant receptors and taste receptors belong to a specific sub-class of GPCRs, with seven trans-membrane helices, and reside in the membranes of olfactory neurons, interestingly where one neuron expresses just one specific receptor (Mori and Sakano, 2011). It has been reported that olfactory receptor genes are diverse and rapidly evolving (Buck and Axel, 1991). For example, Mainland et al. (Mainland et al., 2014) found that 63% of the odorant receptors had genetic polymorphisms that likely altered *in vitro* receptor functions. On average, 30% of odorant receptor alleles were functionally different in two individuals. Residues likely altering functions were found all over the protein (i.e. not restricted to any domain). They also computed the odds that a residue changed function in their arrays did not associate with evolutionary conservation, but deviated from neutral evolution (Mainland et al., 2014). Hoover et al. identified putative evolutionary adaptive trends of smell perception by analyzing the geographic distribution of variants in OR7D4 among 2224 individuals from 43 populations. They concluded that the enormous functional variability among human sense of smell might be ascribed to the specific population selective pressure to local environments or diet (Hoover et al., 2015). Although the low Ka/Ks ratio showed that there were more synonymous than nonsynonymous substitutions, the high frequency of variants and high fixation scores for the deleterious alleles suggested positive selection (Hoover et al., 2015). Another explanation could be attributed relaxed functional constraints (for example in the specific example of OR7D4), which would result in the accumulation of mutations on odorant receptor genes, ultimately, leading to either loss of function or the emergence of new functions over time (Mainland et al., 2014).

Taste receptors are also GPCRs, which are responsible for five main tastes (i.e., sour, sweet, bitter, salty, and umami). There are numerous differences in taste sensation across human populations, owing to rapid evolution and high polymorphism in taste receptors (Foster et al., 2014). Out of the five tastes, interestingly, 24 human *T2R* (taste 2 receptor) genes detecting the bitter taste have been found in our study, significantly more than the genes perceiving other tastes. This could be explained by the importance of the bitter perception which enables human to better avoid potentially harmful or toxic substances (Montmayeur and Matsunami, 2002; Shi et al., 2003). Furthermore, variations in human T2R genes are more prevalent in humans than in other mammals, suggesting that the relaxed selective constraint for the T2R genes has been working in human populations. Supporting this, it has been shown that the efficiency of perceiving bitter taste has been weakened in the human lineage (Niimura and Nei, 2006). In addition, the high-frequency genetic polymorphisms in odorant and taste receptors in the human genome are ascribed to their cooperation in flavor perception, suggesting they have evolved together in vertebrates (Hasin-Brumshtein et al., 2009).

Immune system genes such as HLAs (human leukocyte antigens), KIRs (Killer-cell Ig-like receptors), and TLRs (Toll-like receptors) are ranked among the most variable in the genome (Trowsdale and Parham, 2004). Most of the variations are associated with binding of peptides from pathogens to the grooves of MHC (major histocompatibility complex) I and II molecules (Trowsdale and Parham, 2004). Polymorphism is exploited by those genes in order to conflict with a large number of different pathogens. The wide variation is most likely a result of a selection for highly increasing the defense potential and therefore reduces the risk that the whole population will be eliminated by a pathogen (Davis, 2014).

Drug metabolizing enzymes (DMEs) metabolize a wide range of drugs and naturally generating toxins in the environment. The evolution of DMEs occurred after the animals moved to the land and encountered the plants there (Nebert, 1997). The new environmental chemicals might be responsible for shaping this high diversity of DMEs (Nebert, 1997). For example, the DME genes produced polymorphic enzymes that can enable or diminish the responses to plant diet (Ingelman-Sundberg et al., 1999). SNPs of enzyme metabolizing novel environmental chemicals caused changes in function of the enzyme to give an individual a positive or negative evolutionary selection (Janha et al., 2014). For example if the cytochrome P450 enzyme eliminated the toxic chemical, a gain-of-function allele might be positively selected. By contrast if the enzyme transformed a non-toxic precursor into a toxin, an allele with poor enzyme activity might be selected (Janha et al., 2014).

4.4.2 Variants at the chemical binding sites

Interestingly, many of these variants are at sites predicted or known to interact with chemicals. Such variants are likely to induce direct functional effects on these proteins. For example, the Y402H mutation in CFH/CFHR proteins was identified to increase the risk of age-related macular degeneration by altering the binding affinity to heparin sulphate (HS), a polysaccharide present on the cell surface (Langford-Smith et al., 2014). Several SNPs occurring within a distance of 5Å from their chemical ligands that might result from perturbations were identified by Mechismo system (**Table 4.2**).

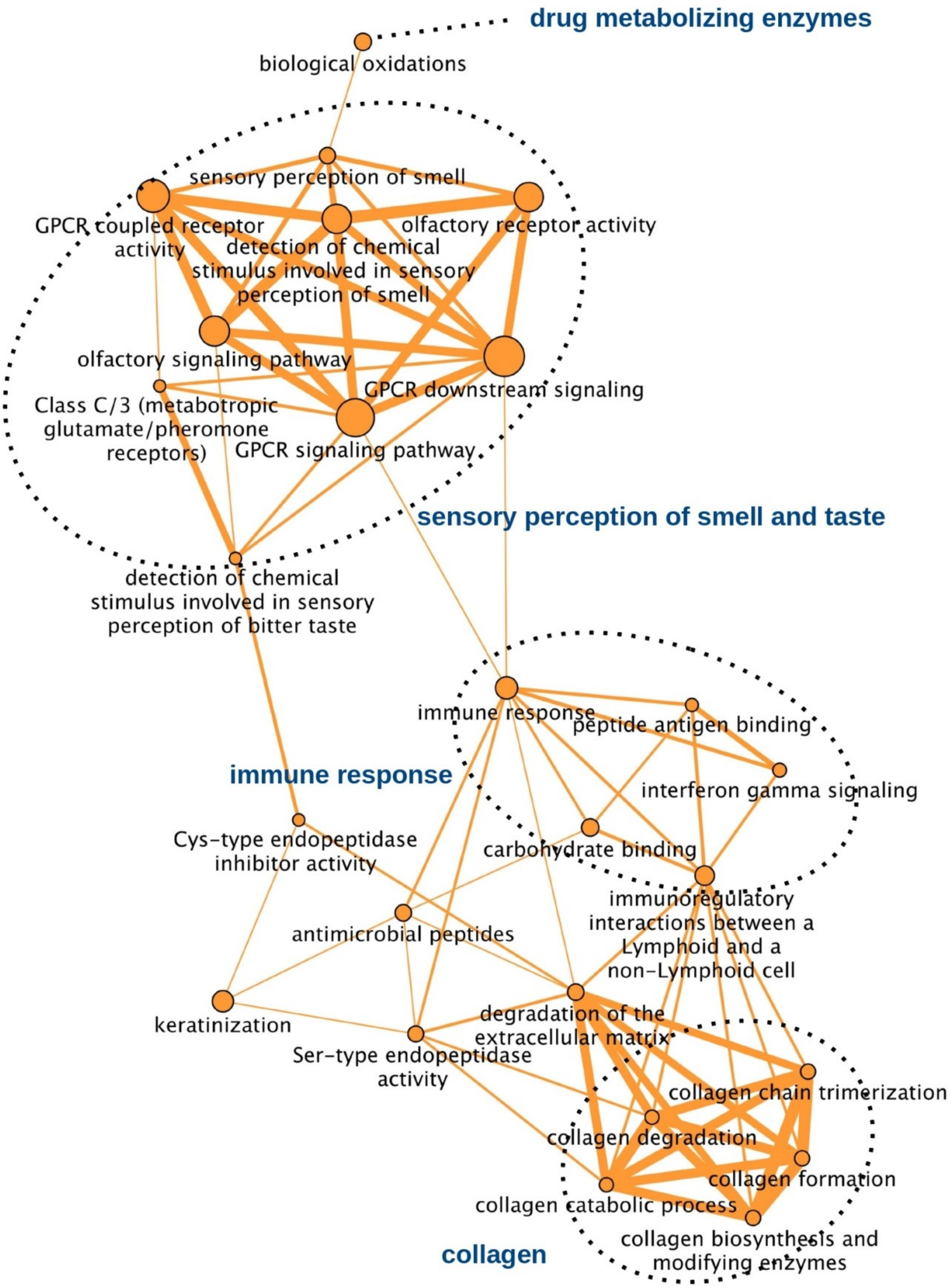

drug metabolizing enzymes
biological oxidations
sensory perception of smell
GPCR coupled receptor activity
olfactory receptor activity
detection of chemical stimulus involved in sensory perception of smell
olfactory signaling pathway
Class C/3 (metabotropic glutamate/pheromone receptors)
GPCR downstream signaling
GPCR signaling pathway
sensory perception of smell and taste
detection of chemical stimulus involved in sensory perception of bitter taste
immune response
peptide antigen binding
immune response
interferon gamma signaling
Cys-type endopeptidase inhibitor activity
carbohydrate binding
immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
antimicrobial peptides
keratinization
degradation of the extracellular matrix
Ser-type endopeptidase activity
collagen chain trimerization
collagen degradation
collagen formation
collagen catabolic process
collagen biosynthesis and modifying enzymes
collagen

Figure 4.1 Gene function enrichment map for genes with more variations. Node size represents the number of genes with many variations within the enriched gene-set. Edge size represents the number of genes that overlap between two gene-sets connected. Highly redundant gene-sets are grouped together as clusters, and groups of functional related gene-sets are manually circled.

Table 4.2 Variants at the chemical binding sites

Uniprot entry/variant	Protein	Site	Chemical
Q31610/L133F	HLA-B	L133F	fucose
Q31610/E176V	HLA-B	E176V	fucose
Q29974/Q99E	HLA-DRB1	Q99E	fucose
Q29974/V114A	HLA-DRB1	V114A	fucose
O60449/E331D	LY75	E331D	fucose
P04439/A174V	HLA-A	A174V	2-amino-4-ethyl sulfanyl butyric acid
P01906/R87T	HLA-DQA2	R87T	acetamidomethylcysteine
P15812/L182Q	CD1E	L182Q	N-Glycoloylganglioside GM2
P24071/D209N	FCAR	D209N	Alpha-GalCer (C20:2)
P04439/H175R	HLA-A	H175R	N-hexacosanoylisoglobotriaosyl ceramide
P30480/L133F	HLA-B	L133F	C12:0-di-sulfatide
P04439/G103R	HLA-A	G103R	glucose monomycolate
P23141/Y366C P23141/M361V P23141/A93V	CES1	Y366C M361V A93V	tamoxifen
P09210/R13Q P09210/D42G P09210/V55L P09210/P110S P09210/S112T	GSTA2	R13Q D42G V55L P110S S112T	etacrynic acid

Several variants at the binding sites of ligands and human MHC I and MHC II were predicted by Mechismo (**Table 4.2**). A study suggested that small carbohydrates (e.g. fucose) attach to the linkers (lipids, e.g. acetamidomethylcysteine, 2-amino-4-ethyl sulfanyl butyric acid) to fulfill the requirements of α/β and γ/δ TCR (T-cell receptor) antigen binding, as part of glycopeptide and carbohydrate-specific T cell responses (Speir et al., 1999).

The four long chain glycolipids, N-Glycoloylganglioside GM2, Alpha-GalCer (C20:2), N-hexacosanoylisoglobotriaosyl ceramide and C12:0-di-sulfatide are α-Galactosylceramide (α-GalCer)-like antigens of type I natural killer T (NKT) cells which initiate immune responses during disease by modulating cell signal transduction events (Fujii et al., 2003). Most NKT cells are activated by CD1d-glycolipid complex through their CDR2β and CDR3α loops which recognize glycolipid antigens (López-Sagaseta et al., 2012). The other human MHC class I-like protein CD1b shows the capability of accommodating broader range of differences in lengths (up to 80 carbons) of acyl chains of glycolipids. For example, the structure of CD1b complex reported in Batuwangala's study presented longer acyl chain lipids (e.g., glucose monomycolate) occupying the whole part of interlinked channels (Batuwangala et al., 2004).

SNPs at the drug binding sites are often essential determinants of inter individual differences in drug response (Langford-Smith et al., 2014), and we identified several of these. For example, the xenobiotic processing and metabolizing enzyme carboxylesterase 1 (CES1) was found to bind to the breast cancer drug tamoxifen (Mésange et al., 2002). In Fleming's drug-CES1 complex structure, tamoxifen bound with eight active site residues within the catalytic pocket via hydrophobic contacts (Fleming et al., 2005). The triphenyl moiety of tamoxifen fits well in the pocket, with two of the rings connecting with Leu388 and Phe101 (Fleming et al., 2005). According to mCSM-lig, the mutations A93V, M361V, Y366C at the active site cavity are predicted to lead to a destabilizing in affinity for tamoxifen and the human CES1 ((PDB ID: 1YA4)) -0.117 log(fold change), -0.518 log(affinity fold change) and -0.736 log(affinity fold change), respectively.

Ethacrynic acid (EA) is a potent loop diuretic drug used to lower high blood pressure by inhibiting glutathione S-transferases (GSTs) (Somberg and Molnar, 2009). It is a phenoxyacetic acid derivative with a ketone moiety and a methylene group in the side chain. GSTs catalyze the reaction of the methylene group with glutathione (GSH) to form the active form - glutathione conjugate (Ploemen et al., 1993). The complex structures of alpha-class GSTA1-1 with EA and EA-GSH have been reported (Cameron et al., 1995). The EA moiety binds with GST via the dichlorophenoxy group in van der Waals contact with Gly14 and Phe10, Leu107, Leu108, Val111, Met208 and Phe222. The carboxylate oxygens contact with three hydrogen interactions, one with Gly14 and two with water molecules. One has hydrogen bonds to the amino nitrogen of Arg13 and the carbonyl oxygen of Pro207. In the structure of EA-GSH complex, one of the carboxylate oxygens has two hydrogen bonds, one with the carbonyl oxygen of Val55 and the other one with the amino nitrogen of Val55 through a water molecule (Cameron et al., 1995). Site-directed mutants of R13Q, V55L, P110S and S112T are within bonding distance may have effects on the catalytic properties of the enzyme. In mCSM-lig, V55L and P110S are predicted to destabilized by -0.715 log(affinity fold change) and -0.047 log(affinity fold change), separately.

4.4.3 Influences of variations on toxicity responses

Genetic variations in drug metabolizing genes could also lead to toxicity and thereby to drug withdrawal or discontinuation in specific populations. We extracted interactions of eight withdrawn drugs with their targets and metabolizing enzymes (or homologous proteins) from Mechismo to predict mechanistic consequences (e.g., toxicity) for variations (**Figure 4.2**).

Bromoergocryptine was one of the ergot alkaloids acting as the dopamine receptor agonist for treating prolactin secretion, type II diabetes, and Parkinson disease. *In vivo* experiments have found that bromoergocryptine is primarily oxidized by CYP3A4 near the heme iron with its tripeptide moiety. Two particular residues, Arg212 and Thr224, have been observed to sustain their binding conformation where Arg212 directs bromoergocryptine to approach to the heme and Thr224 binds to N1 atom of bromoergocryptine by an H-bond (Sevrioukova and Poulos, 2012). By searching Mechismo, an amino acid substitution, R246H in CYP4F22 with 26.53% sequence similarity to CYP3A4 was found at the interface that might influence the drug response. It was reported that bromoergocryptine was withdrawn from the market in the USA due to its neurological toxicity (Liu et al., 2013).

Troglitazone was one of the bile salt export pump potent inhibitors, but troglitazone-induced hepatocellular injury led to its withdrawal from the market. It contained the structural alert 4-ethylphenol which could be activated to form adducts with GSH and induced GSH depletion in hepatocytes (Lang et al., 2007). The co-crystallized structure of troglitazone-CYP2C8 showed that the chromane ring approached to the heme iron, and the thiazolidinedione group was closer to the C’ end of chain B. Arg241 on helix G and Asn204 on helix F donated two hydrogen bonds to the two carbonyl oxygens of the thiazolidinedione ring, and Val296 and Glu300 on helix I donated two hydrogen bonds to the phenolic oxygen. The ether oxygen of 4-ethyphonel at the middle part of troglitazone was hydrogen-bonded to Ser103 on helix B (Schoch et al., 2008). Mechismo identified a substitution T298M (in 0.7% of South Asians) at the interface of troglitazone and CYP2W1 (with 24.26% sequence similarity to CYP2C8) which might impact catalytic activity.

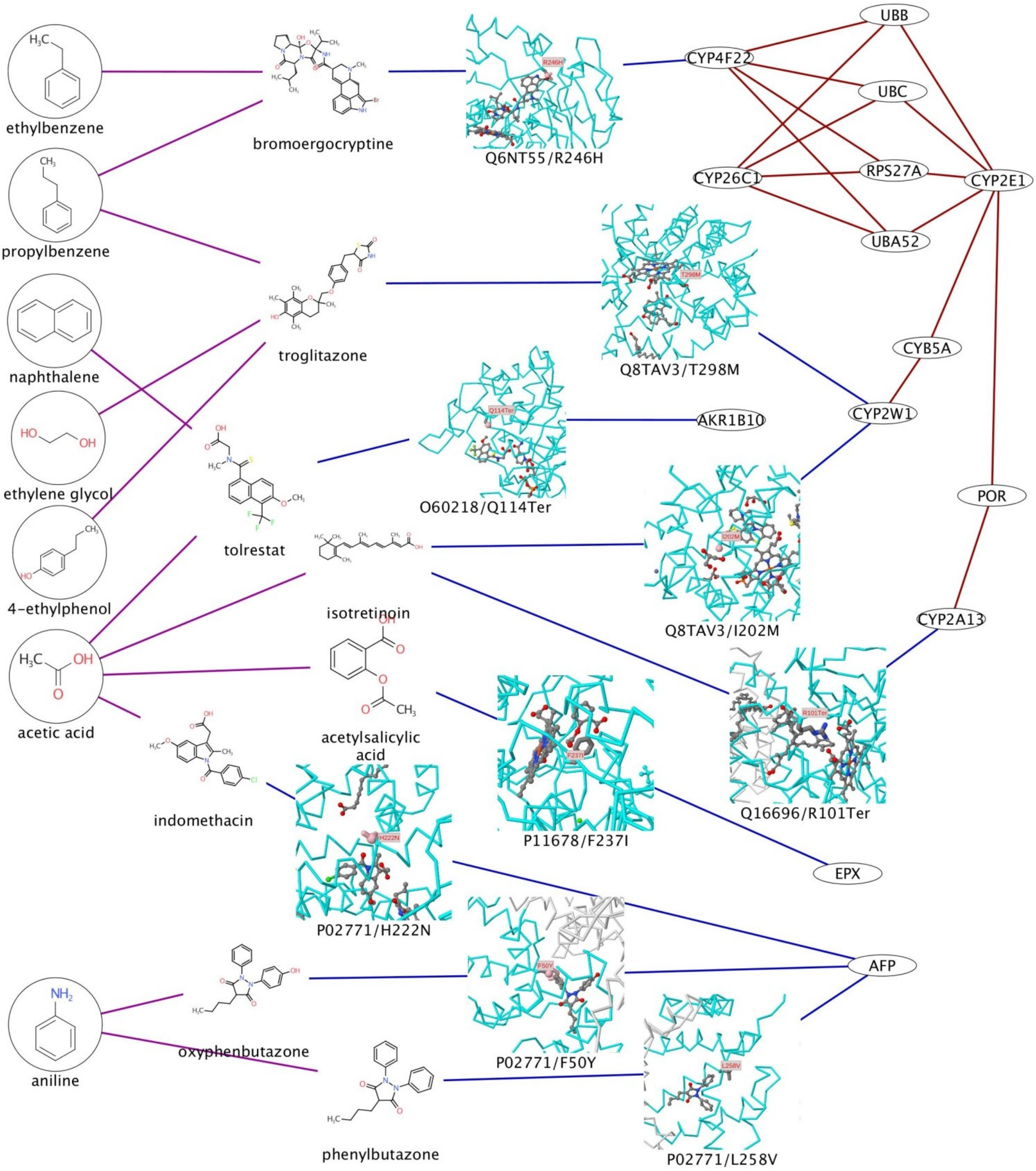


Figure 4.2 Interactions of eight withdrawn drugs with their structural alerts and targets/metabolizing enzymes present how SNPs on the interface impact the binding activity, leading to drug toxicity

Another substrate of CYP2C8, isotretinoin, was prescribed as severe acne treatment, binding to its enzyme by hydrogen bonds between acetic acid moiety and Asn204 and Arg241 (Schoch et al., 2008). The movement of Arg241 from outside of the cavity to inside offered a strong and charge-stabilized hydrogen bond with the acetic acid of isotretinoin. The changes of I202M, with a frequency of 0.1% in European people and none in other populations, was seen in Mechismo might impact the binding event. Isotretinoin was withdrawn from the Italian market due to its effect on male reproduction (Alli and Yorulmaz, 2015).

Tolrestat is one of the aldose reductase inhibitors which was approved for the control of diabetes but was withdrawn from the market due to its toxicity. Tolrestat bound crystal structure of AKR1B10 showed that the catalytic sites contained Tyr49, His111, NADP+, and Trp112, of which Tyr49 and Trp112 donated hydrogen bonds to acetic acid group of tolrestat. It was observed from Mechismo that near the active site, there was a substitution Q114Ter which could affect the conformational changes of AKR1B10 with ligand binding. This might result in the formation of induced cavity in AKR1B10, leading to a broad ligand specificity (Shah et al., 2012).

Acetylsalicylic acid (aspirin) is a nonsteroidal anti-inflammatory drug, but was withdrawn from British marketplace due to hepatic and neurological toxicities (Aronson, 2004). Acetylsalicylic acid can be oxidized to acetylsalicylic acid free radical by peroxidase enzyme (e.g., eosinophil peroxidase (EPX)) (Singh et al., 2010). Acetylsalicylic acid bound to EPX by forming one hydrogen bond with a heme iron, which in turn generated two additional hydrogen bonds with Gln105 and His109. Two atoms of the acetic acid moiety formed other hydrogen bonds with Phe422 and Gln423. It is proposed that acetylsalicylic acid is balanced favorably at the active site of the peroxidase (Singh et al., 2010). Compared with other EPX inhibitors, the interactions of Phe422 and Gln423 with the acetic acid are unique, and suggest possible insights into the mechanism of acetylsalicylic acid toxicity (Singh et al., 2010). The substitution F237I (PDB: 2qqt) found in Mechismo could interrupt the interaction of acetylsalicylic acid and EPX.

Indometacin is also a nonsteroidal anti-inflammatory agent used to reduce pain, fever and swelling from inflammation and was withdrawn in Germany due to gastrointestinal toxicity. The complex of indometacin bound human serum albumin (HSA) showed that the acetic acid group binded to Arg218 by a salt bridge, and chlorobenzoyl group lied at the bottom of the cavity, thereby the substitution of H222N (approximately 0.1% of the European populations) found in the binding sites of alpha-fetoprotein (AFP, 41.13% similarity to HSA) might generate protein causing drug binding defects (Ghuman et al., 2005).

Oxyphenbutazone and phenylbutazone are potent nonsteroidal anti-inflammatory drugs but are associated with liver damage (Kirkland and Fowler, 2010). The potential liver toxicity mechanism was illustrated in two studies (Stepan et al., 2011; Kim et al., 2016). Specifically, most drugs containing structural alert aniline can undergo metabolic activation leading to the generation of reactive oxygen species. These activate the antioxidant response element and PPAR gamma signaling pathways, and disrupt human tyrosyl-DNA phosphodiesterase 1 and thyroid receptor signaling pathways (Kim et al., 2016). Oxyphenbutazone and phenylbutazone both interacted with HSA via a single hydrogen bond with Tyr150. The additional hydroxyl group near the mouth of the pocket possessed by oxyphenbutazone leads to a rotation of itself by 180 degrees compared with phenylbutazone. The phenolic hydroxyl of oxyphenbutazone interacts with Arg257 and the modification F50Y on the interface, highlighted by Mechismo, might have an effect drug binding. The phenyl moiety of phenylbutazone also made a hydrogen bond with Arg257 and the modification L258V next to it might modify the affinity of drug binding (Ghuman et al., 2005). The frequencies are about 3.7% in South Asians, and 2.0% and 2.1% in Americans and Europeans, and only 0.1% in East Asians. Oxyphenbutazone has been withdrawn in three south Asian countries, seven European countries, and the USA.

4.5 Conclusion and discussion

Much of the enthusiasm advances in "-omics" have focused on their application to understanding human health and diseases. In this chapter, we investigated the distribution of single nucleotide polymorphisms in worldwide human population samples measured by 1000 Genomes Project. Remarkably, some suspicious categories of genes related to environmental adaptation, such as those involved in sensory perception functions (e.g., odorant and taste receptors) and immunological functions and drug responses (e.g., action and toxicity) were found to be significantly enriched for nonsynonymous SNPs, hinting that those seem to be targets of human selection processes. Based on the observation that nonsynonymous SNPs in the ligand binding domain could change the ligand binding affinity to the receptor, we systematically investigated the effects of SNP variations in the ligand binding sites on biological processes. Identifying the relationships of genetic variations and chemical molecules might provide insights into the mechanisms of how SNPs affect sensory perception, diet metabolism, drug response and common diseases, such as hypertension, diabetes, and cancer. This study shows the potential of predicting inter-individual variability in drug toxicity response by integrating data from multiple datasets. A similar prediction was done in an open community challenge, where multiple groups developed algorithms for predicting the effects of toxic chemicals on different individuals based on genetic profiles and the structural attributes of chemicals (Eduati et al., 2015). Though initially promising, improving the accuracy of the predictions would require very large sample sizes, because the individual SNPs represented only a small part of overall variations in toxic response (Chatterje et al., 2013).

Nevertheless, pharmacogenetics could provide easier and direct diagnostic applications by identifying the association between genotype and drug response, for example, to screen the drug with highest efficacy among several candidates or to avoid severe adverse drug reactions (Roses, 2002). It is noteworthy that pharmacogenetics prescribing is limited for a few drugs due to the problems in identifying replicated associations of phenotype and drug reaction, ethical and regulatory considerations (Lauschke 2016). Conquering these problems is of crucial importance to further personalize medicine, which could help to deduce morbidity and mortality for patients and provide a more rational distribution of health-care resources.

Bringing about the health-care benefits of pharmacogenetics requires a close corporation between researchers and clinicians. However, so far only phenotyping for thiopurine methyltransferase (TPMT) has been substantially put into the clinical practical to prospectively identify TPMT-deficient and heterozygous TPMT patients (Relling et al., 2011). Approximately 3-14% people of the population have heterozygous TPMT genotype and possess low enzyme activity. These patients are likely to experience severe myelosuppression with thiopurine drugs treatment, such as azathioprine, and thioguanine (Relling et al., 2011). Ideally, researchers and clinicians should be involved together in pharmacogenetics research involving patients to boost health-care.

Overall, our study tested the ability of computational methods to provide understandable mechanistic insights about individual differences in sensory perception, diet, and drug response by using genetic variation data. Although the suggested methods would require further actions to verify, we hope that the analytics presented in this chapter would spark interest to other scientists and regulators to explore the data for the benefit of the entire community.

CHAPTER V: General discussion and conclusions

This dissertation presented a combined systems biology and chemical approach to assess the potential toxicities of drugs based on their chemical structures and biological features. The proposed framework integrated chemical space and pharmacological space to extract the associations of chemical fragments and toxicities, and predicted the potential toxic effects of unknown drugs based on the structural alerts (Chapter II). In efforts to offer biological interpretability of the outputs of the proposed approach and minimize the false positive of the predictions, mechanisms of toxicity have been investigated with gene expression data and pathways/networks that were perturbed by toxins (Chapter III). Additionally, the relationships of genetic variants and phenotype were extensively analyzed to provide insights into how genetic polymorphisms can be related to differences in drug response among individuals (Chapter IV).

5.1 Summary of major findings

5.1.1 Data-driven identification of structural alerts for predicting drug toxicity

We developed a workflow that derived chemical substructures/fragments and toxicity information from current available public big data sources, and performed statistical analyses and machine learning methods to explore the fragment-toxicity relationships, and to identify substructures/fragments (structural alerts) that associated with toxic effects. Toxicity prediction models, using both traditional SVM and SCCA computational approaches, based on those structural alerts showed good overall accuracy in addition to a capability to extract associations between chemical fragments and toxicities. Predictions conducted for withdrawn drugs confirmed the practicability of the structural alerts dataset and computational methods we proposed. Overall, this workflow reduces a large potential chemical space to smaller dataset of structural alerts that can be manually curated by human experts to provide potentially more accurate and effective toxic predictions for unknown compounds.

5.1.2 Gene expression analysis provided mechanisms of drug toxicity

Toxicogenomics are valid assays to clarify the molecular pathways involved in drug activity. In chapter III, we found that toxicogenomics assays tended to more reflect the mechanism of drug action rather than its underlying toxicity. It seemed unanticipated since the target tissues (e.g., liver and kidney) were not traditionally considered as targets of many drugs (e.g., acetaminophen). Nevertheless, many known and potentially novel biomarkers are evident in the toxicogenomics data, which further validates the power of toxicogenomics. Overall, the findings help interpret these toxicity prediction methods, suggesting a necessity to include (or exclude) mechanistic contributions to gene expression profiles before extracting (for example) multi-gene predictors. Furthermore, we found *in vitro* gene expression arrays were almost always in poor agreement with *in vivo* equivalents when considering data on large set of compounds, which argues for caution in the use of these approaches as a replacement for *in vivo* methods.

5.1.3 Genetic polymorphisms analysis provided mechanisms of individual differences of drug toxicity

We expanded the use of transcriptomic profiles to genetic profiles to help understand mechanisms underlying genetic variation in toxic response on the population level. These variants are often found in drug ADME genes, likely often changing the enzymatic activity of the gene products, which in turn could result in altered drug toxicity or efficacy. In addition, the higher frequency of variants was found in MHC genes, indicated that genetic variants could also modulate the risk of immune-mediated toxicity. Moreover, we found some chemical compounds containing the same scaffold were in conjunction with specific genotype associated with certain susceptibility, e.g., diabetes, suggesting that the combined chemical structural and genetic effect may contribute to the underlying risk factors.

5.2 Contributions and practical implications

5.2.1 Structural alert concept as applied in drug discovery

The proposed approach of fragment-based toxicity prediction is thought to be practical at different stages of the drug discovery process (and related subjects). At the early stage, the structural alert concept would be advisable to screen the drug candidates that could might otherwise proceed further to the later stages at great cost. The approach could be also used in drug repurposing process to find new indications (i.e., new diseases) for existing drugs by detecting their potentially beneficial side effects. A well-known example is thalidomide, launched for treating morning sickness and later was withdrawn due to its teratogenicity, but then later redeveloped as a treatment for multiple myeloma (Ashburn and Thor, 2004). The structural alerts could also be exploited in fragment-based drug discovery approaches where can filter toxicity-related fragments from the chemical space of drug-like molecules fragments (Siegel and Vieth, 2007).

5.2.2 Integrative gene expression-chemical features for toxicity assessment

Large scale analysis of gene expression profiles is a valuable tool for explaining mechanisms for chemotherapeutic drugs and predicting toxicity of unknown compounds. Apparently, different compounds generate particular expression profile and similarities of expression profiles among compounds can demonstrate similarities of their therapeutic and toxic mechanisms. We have found that a systematic approach, combined with information on biological and chemical features, is probably required to improve the accuracy of predictions of drug toxicity based on expression profiles. The challenge is to connect chemical features with biological responses. Our study in Chapter III confirmed the value of a chemical-centric perspective of phenotypes to identify the effects of chemicals on phenotypic outcomes as mentioned before (Duran et al., 2014) and will help to relate chemical features to systemic responses.

5.2.3 Pharmacogenomic characterization of drug response enables personalized medicine

The integrative genetic and systems biology analysis provides great potential for understanding the associations of genetic variants and drug responses (e.g., toxicity) as well as for personalization of treatments. The assessment of the changes in protein-ligand affinity upon variation may improve our understanding of binding mode and help us to choose more efficacious therapeutics. Genetic studies of single nucleotide polymorphisms in various gene classes across compounds and across individuals provide a potential strategy for real-world applications (e.g., population-specific sensory perception and diet). Chemical structural feature/genotype/phenotype can be considered as complementary patient-specific parameters to classical factors (e.g., age) to assess their susceptibility to the specific treatment response so as to provide clinical decision support to improve patient care.

5.3 Limitations and future solutions

5.3.1 Chemical data limitations and future solutions

The proposed method of systematically screening structural alerts highly depends on the pre-defined chemical fragments and toxicity terms. Future work could test the performance with other types of fingerprints and toxicity information. For example, the ECFP4 fingerprint provided by ChemAxon (http://www.chemaxon.com) as well as chemical structure descriptors supplied by commercial software Dragon (https://chm.kode-solutions.net/products_dragon.php) and toxicity keywords in PharmaPendium (https://www.pharmapendium.com/). Another way of getting chemical substructures is by directly fragmenting graph structures of chemical compounds into fragments based on the recap-rules (Lewell et al., 1998) and the principles of graph theory and depth-first algorithms.

5.3.2 Biological data limitations and future solutions

The toxicogenomic signatures used in this study were from *in vitro*, *in vivo* rodent and *in vivo* human models. The potential limitation is how to bridge species differences and the *in vivo-in vitro* gap in terms of drug efficacy. For that, significant validation is required to examine drug ADME processes. Another limitation is that patient-related factors contributing to the efficacy of drugs are not currently well understood. To consider these individual contributing factors, future work must incorporate larger sample sizes, which argues for large, multi-disciplinary national and international collaborations. In parallel, GWAS and large-scale genome sequencing data will help to characterize the associations of phenotypic outcomes and gene mutations.

5.3.3 Methodological limitations and future solutions

Structural alert-based prediction models currently have comparatively poor results when predicting toxicities. The best results rely on data normalization, which requires practical tests of the methods. The ideal models should integrate all potential mechanisms although currently, their performances suffer when predicting multiple toxic endpoints with complex mechanisms. Methods are also limited in detecting changes in protein-small molecules stability upon a genetic variation. Hence, a universal predictive tool including structure signatures related to changes in affinity (e.g determined by modelling the binding sites) needs to be developed or improved.

5.4 Epilogue

Integrating chemical features with biological features, including gene expression, genetic data, and proteomics/metabonomics, is one of the next great challenges faced by biomedicine. Our studies assessing chemical toxic outcomes at systems and population levels in this thesis will help pave the way for new work on toxicology and disease. Methods like those presented in this thesis will be essential in developing new medicines and improving diagnostics in the new era of personalized medicine.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A. et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods. 7*, 248-249.

Afzal, Z., Schuemie, M.J., van Blijderveen, J.C., Sen, E.F., Sturkenboom, M.C. et al. (2013). Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC Med. Inform. Decis. Mak.* 13, 30.

Agresti, A. (2002). Categorical data analysis. *Second edition. New York: Wiley.* 91-101.

Ahmed, R., Branley, H.M. (2009). Reversible bronchospasm with the cardio-selective beta-blocker celiprolol in a non-asthmatic subject. *Respir. Med. CME 2*, 141-143.

Ahmed, J., Worth, C.L., Thaben, P., Matzig, C., Blasse, C. et al. (2011). FragmentStore - a comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic. Acids Rec. 39*, D1049-D1054.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V. et al. (2016). The Ensembl gene annotation system. *Database (Oxford)*, 2016. pii: baw093.

Allen, J.R., Hargraves, W.A., Hsia, M.T.S., Lin, F.S.D. (1979). Comparative toxicology of chlorinated compounds on mammalian species. *Pharmacol. Ther. 7*, 513-547.

Alli, N., Yorulmaz, A. (2015). An unusual side effect of isotretinoin: retinoid dermatitis affecting external urethral meatus. *Gutan. Ocul. Toxicol. 34*, 176-177.

An, Y.R., Kim, J.Y., Kim, Y.S. (2016). Construction of a predictive model for evaluating multiple organ toxicity. *Mol. Cell. Toxicol. 12*, 1-6.

Ankley, G.T., Bennett, R.S., Erickson, R.J., Hoff, D.J., Hornung, M.W. et al. (2010). Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem. 29*, 730−741.

Antczak, P., White, T.A., Giri, A., Michelangeli, F., Viant, M.R. et al. (2015). Systems biology approach reveals a calcium-dependent mechanism for basal toxicity in Daphnia magna. *Environ. Sci. Technol.* 49, 11132-11140.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S., Ceol, A. et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods 8*, 528-529.

Aronson, J. (2004). Stephens' Detection of New Adverse Drug Reactions, 5th edn. *Br. J. Clin. Pharmacol. 58*, 227.

Ashburn, T.T., Thor, K.B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov. 3*, 673-683.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H. et al. (2000). *Nat. Genet. 25*, 25-29.

Ashby, J., Tennant, R.W. (1988). Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutat. Res. 204*, 17-115.

Ashby, J., Tennant, R.W. (1991). Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. *Mutat.*

Res. 257, 229-306.

Afshari, C.A., Hamadeh, H.K., Bushel, P.R. (2011). The evolution of bioinformatics in toxicology: Advancing toxicogenomics. *Toxico. Sci. 120*, S225-S237.

Bader, G.D., Betel, D., Hogue, C.W.V. (2003). BIND: the biomolecular interaction network database. *Nucleic. Acids Res. 31*, 248-250.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics 16*, 412-424.

Barh, D., Zambare, V., Azevedo, V. (2013). *OMICS: applications in biomedical, agricultural, and environmental sciences.* CRC Press.

Batuwangala, T., Shepherd, D., Gadola, S.D., Gibson, K.J.C., Zaccai, N.R. et al. (2004). The crystal structure of human CD1b with a bound bacterial glycolipid. *J. Immunol. 172*, 2382-2388.

Bender, A., Mussa, H.Y., Glen, R.C., Reiling, S. (2004). Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci. 44*, 170-178.

Benigni, R. (1991). QSAR prediction of rodent carcinogenity for a set of chemicals currently bioassayed by the US National Toxicology Program. *Mutagenesis 6*, 423-425.

Benigni, R., Bossa, C. (2008). Structural alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat. Res. Rev. 659*, 248-261.

Benigni, R., Bossa, C. (2011). Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. *Chem. Rev. 111*, 2507-2536.

Bertilsson, G., Heidrich, J., Svensson, K., Asman, M., Jendeberg, L. et al. (1998). Identification of a human nuclear receptor defines a new signaling pathway for CYP3A induction. *Proc. Natl. Acad. Sci. U. S. A. 95*, 12208-12213.

Betts, M.J., Lu, Q., Jiang, Y.Y., Drusko, A., Wichmann, A. et al. (2015). Mechismo: predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucl. Acids Res. 43*, e10.

Blower, P.E., Yang, C., Fligner, M.A. Verducci, J.S., Yu, L. et al. (2002). Pharmacogenomic analysis: correlating molecular substructure classes with microarray gene expression data. *Pharmacogenomics J. 2*, 259-271.

Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H. (2008). Chapter 12 – PubChem: Integrated platform of small molecules and biological activities, in: *Annual Reports in Computational Chemistry*, 217-241.

Bonzo, J. A., Rose, K., Freeman, K., Deibert, E., Amaral, K. B. et al. (2015). Differential effects of trovafloxacin on TNF-α and IL-6 profiles in a rat hepatocyte–Kupffer cell coculture system. *Applied In Vitro Toxicology 1*, 45-54.

Bowes, J., Brown, A.J., Hamon, J., Jarolimek, W., Sridhar, A. et al. (2012). Reducing safety-related drug attrition: the use of *in vitro* pharmacological profiling. *Nat. Rev. Drug Discov. 11*, 909-922.

Brink, R. H., Walker, J. D. (1987) EPA TSCA ITC Interim Report, *Dynamic Corporation,* *Rockville*

Buck, L., Axel, R. (1991). A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell 65*, 175-187.

Byrns, M.C., Predecki, D.P., Peterson, L.A. (2002). Characterization of nucleoside adducts of cis-2-butene-1,4-dial, a reactive metabolite of furan. *Chem. Res. Toxicol. 15*, 373-379.

Carbonell, P., Lopez, O., Amberg, A., Pastor, M., Sanz, F. (2017). Hepatotoxcity prediction by systems biology modeling of disturbed metabolic pathways using gene expression data. *ALTEX 34*, 219-234.

Caiment, F., Tsamou, M., Jennen, D., Kleinjans, J. (2014). Assessing compound carcinogenicity *in vitro* using connectivity mapping. *Carcinogenesis 35*, 201-207.

Cameron, A.D., Sinning, I., L'Hermite, G., Olin, B., Board, P.G. et al. (1995). Structural analysis of human alpha-class glutathione transferase A1-1 in the apo-form and in complexes with ethacrynic acid and its glutathione conjugate. *Structure 3*, 717-727.

Carabello, B.A. and Crawford, F.A. (1997). Valvular heart disease. *N. Engl. J. Med. 337*, 32-41.

Chatr-aryamontri, A., Breitkreutz, B., Oughtred, R., Boucher, L., Heinicke, S. et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res. 43(Database issue)*, D470-D478.

Chatterje, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J. et al. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet. 45*, 400-405.

Chen, L.J., Hecht, S.S., Peterson, L.A. (1997). Characterization of amino acid and glutathione adducts of cis-2-butene-1,4-dial, a reactive metabolite of furan. *Chem. Res. Toxicol. 10*, 866-874.

Chen, M., Zhang, M., Borlak, J. and Tong, W. (2012). A decade of toxicogenomic research and its contribution to toxicological science. *Toxicol. Sci.* 130, 217-228.

Chen, Q.Y., Alarcon, S., Tharp, A., Ahmed, O.M., Estrella, N.L. et al. (2009). Perceptual variation in umami taste and polymorphisms in TAS1R taste receptor genes 1,2,3,4. *Am. J. Clin. Nutr. 90*, 770S-779S.

Chen, Y., Cunningham, F., Rios, D., McLaren, W.M., Smith, J. et al. (2010). Ensembl variation resources. *BMC Genomics 11*, 293.

Cunningham, A.R., Moss, S.T., Iype, S.A., Qian, G., Qamar, S. et al. (2008). Structure-activity relationship analysis of rat mammary carcinogens. *Chem. Res. Toxicol. 21*, 1970-1982.

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B. et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res. 33*, e175.

Dang, N.L., Hughes, T.B., Miller, G.P., Swamidass, S.J. (2017). Computational Approach to Structural Alerts: Furans, Phenols, Nitroaromatics, and Thiophenes. *Chem. Res. Toxicol. 30*, 1046-1059.

Davis, D.M. (2014). The compatibility gene. How our bodies fight disease, attract others, and define ourselves. Oxford: Oxford University Press. ISBN 0-19-931641-4.

Duran-Frigola, M., Rossell, D., Aloy, P. (2014). A chemo-centric view of human health and disease. *Nat. Commun. 5*, 5676.

Eapen, S., Singh, S., D'Souza, S.F. (2007). Advances in development of transgenic plants for remediation of xenobiotic pollutants. *Biotechnol. Adv. 25*, 442-451.

Eduati, F., Mangravite, L.M., Wang, T., Tang, H., Bare, J.C. et al. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol. 33*, 933-940.

Ekins, S. (2016). The next era: Deep learning in pharmaceutical research. *Pharm. Res. 33*, 2594-2603.

Evans, W.E. (2004). Pharmacogenetics of thiopurine S-methyltransferase and thiopurine therapy. *Ther. Drug Monit. 26*, 186-191.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K. et al. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Res. 44(D1)*, D481-D487.

Fakih, M., Wong, R. (2010). Efficacy of the monoclonal antibody EGFR inhibitors for the treatment of metastatic colorectal cancer. *Curr. Oncol. 17(Suppl 1)*: S3-S17.

Fernald, G.H., Altman, R.B. (2013). Using molecular features of xenobiotics to predict hepatic gene expression response. *J. Chem. Inf. Model. 53*, 2765-2773.

Ferrari, T., Cattaneo, D., Gini, G., Golbamaki, Bakhtyari., N. et al. (2013). Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. *SAR QSAR Environ. Res. 24*, 365-383.

Fey, S.J., Wrzesinski, K. (2012). Determination of drug toxicity using 3D spheroids constructed from an immortal human hepatocyte cell line. *Toxicol. Sci. 127*, 403-411.

Fleming, C.D., Bencharit, S., Edwards, C.C., Hyatt, J.L., Tsurkan, L. et al. (2005). Structural insights into drug processing by human carboxylesterase 1: tamoxifen, mevastatin, and inhibition by benzil. *J. Mol. Biol. 352*, 165-177.

Foster, S.R., Roura, E., Thomas, W.G. (2014). Extrasensory perception: odorant and taste receptors beyond the nose and mouth. *Phamacol. Ther. 142*, 41-61.

Fowler, L.M., Moore, R.B., Foster, J.R., Lock, E.A. (1991). Nephrotoxicity of 4-aminophenol glutathione conjugate. *Hum. Exp. Toxicol. 10*, 451-459.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M. et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res. 41*, D808-D815.

Fujii, S., Shimizu, K., Smith, C., Bonifaz, L., Steinman, R.M. (2003). Activation of natural killer T cells by alpha-galactosylceramide rapidly induces the full maturation of dendritic cells *in vivo* and thereby acts as an adjuvant for combined CD4 and CD8 T cell immunity to a coadministered protein. *J. Exp. Med. 198*, 267-279.

Gao, W., Mizukawa, Y., Nakatsu, N., Minowa, Y., Yamada, H. et al. (2010). Mechanism-based biomarker gene sets for glutathione depletion-related hepatotoxicity in rats. *Toxicol. Appl. Pharmacol. 247*, 211-221.

Geerts, H., Roberts, P., Spiros, A., Potkin, S. (2015). Understanding responder neurobiology in schizophrenia using a quantitative systems pharmacology model:

application to iloperidone. *J. Psychopharmacol. 29*, 372-382.

Ghuman, J., Zunszain, P.A., Petitpas, I., Bhattacharya, A.A., Otagiri, M. et al. (2005). Structural basis of the drug-binding specificity of human serum albumin. *J. Mol. Bio. 353*, 38-52.

Gillard, E.F., Otsu, K., Fujii, J., Khanna, V.K., de Leon, S. et al. (1991). A substitution of cysteine for arginine 614 in the ryanodine receptor is potentially causative of human malignant hyperthermia. *Genomics 11*, 751-755.

Gray, N.S., Wodicka, L., Thunnissen, A.M., Norman, T.C., Kwon, S. et al. (1998). Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science 281*, 533-538.

Gribskov, M., Robinson, N. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem. 20*, 25-33.

Grinberg, M., Stöber, R.M., Edlund, K., Rempel, E., Godoy, P. et al. (2014). Toxicogenomics directory of chemically exposed human hepatocytes. *Arch. Toxicol. 88*, 2261-2287.

Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E. et al. (2014). PDBe: protein data bank in Europe. *Nucleic Acids Res. 42 (Database Issue)*, D285-D291.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn. 46*, 389-422.

Hamadeh, H.K., Amin, R.P., Paules, R.S., Afshari, C.A. (2002). An overview of toxicogenomics. *Curr. Issues Mol. Biol. 4*, 45-56.

Hansch, C., Gao, H. (1997). Comparative QSAR: Radical reactions of benzene derivatives in chemistry and biology. *Chem. Rev. 97*, 2995-3060.

Hansch, C., Leo, A., Mekapati, S.B., Kurup, A. (2004). QSAR and ADME. *Bioorg. Med. Chem. 12*, 3391-3400.

Harmon, G.S., Lam, M.T., Glass, C.K. (2011). PPARs and lipid ligands in inflammation and metabolism. *Chem. Rev. 111*, 6321-6340.

Hartmann, E., Spinweber, C.L. (1979). Sleep induced by L-tryptophan. Effect of dosages within the normal dietary intake. *J. Nerv. Ment. Dis. 167*, 497-499.

Hasin-Brumshtein, Y., Lancet, D., Olender, T. (2009). Human olfaction : from genomic variation to phenotypic diversity. *Trends Genet. 25*, 178-184.

Helma, C., Cramer, T., Kramer, S., De Raedt, L. (2004) Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci. 44*, 1402-1411.

Hewitt, M., Enoch, S.J., Madden, J.C., Przybylak, K.R., Cronin, M.T. (2013). Hepatotoxicity: a scheme for generating chemical categories for read-across, structural alerts and insights into mechanism(s) of action. *Crit. Rev. Toxicol. 43*, 537-558.

Hirode, M., Omura, K., Kiyosawa, N., Uehara, T., Shimuzu, T. et al. (2009). Gene expression profiling in rat liver treated with various hepatotoxic-compounds inducing coagulopathy. *J. Toxicol. Sci. 34*, 281-293.

Hodgson, E., Goldstein, J.A. (2001). Metabolism of toxicants phase I reactions and pharmacogenetics. In *Introduction to biochemical toxicology (67-113). Wiley-Interscience.*

Hoover, K.C., Gokcumen, O., Qureshy, Z., Bruguera, E., Savangsuksa, A. et al. (2015). Global survey of variation in a human olfactory receptor gene reveals signatures of non-neutral evolution. *Chem. Senses 40*, 481-488.

Huang, Q., Dunn, R.T., Jayadev, S., DiSorbo, O., Pack, F.D., Farr, S.B. et al. (2001). Assessment of cisplatin-induced nephrotoxicity by microarray technology. *Toxicol. Sci. 63*, 196-207.

Huang, Y.S. (2007). Genetic polymorphisms of drug-metabolizing enzymes and the susceptibility to antituberculosis drug-induced liver injury. *Expert. Opin. Drug MeTable Toxicol. 3*, 1-8.

Hussaini, S.H., Farrington, E.A. (2007). Idiosyncratic drug-induced liver injury: an overview. *Expert Opin. Drug Saf. 6*, 673-684.

Ichimura, T., Bonventre, J.V., Bailly, V., Wei, H., Hession, C.A. et al. (1998). Kidney injury molecule-1 (KIM-1), a putative epithelial cell adhesion molecule containing a novel immunoglobulin domain, is up-regulated in renal cells after injury. *J. Biol. Chem. 273*, 4135-4142.

Ichimura, T., Asseldonk, E.J., Humphreys, B.D., Gunaratnam, L., Duffield, J.S. et al. (2008). Kidney injury molecule-1 is a phosphatidylserine receptor that confers a phagocytic phenotype on epithelial cells. *J. Clin. Invest. 118*, 1657-1668.

Ingelman-Sundberg, M. (2015). Personalized medicine into the next generation. *J. Intern. Med. 277*, 152-154.

Ingelman-Sundberg, M., Oscarson, M., McLellan, R.A. (1999). Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. *Trends Pharmacol. Sci. 20*, 342-349.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J. et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*, 249-264.

Irwin, J.J., Shoichet, B.K. (2005). ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model. 45*, 177-182.

Iyanagi, T. (2007). Molecular mechanism of phase I and phase II drug metabolizing enzymes: implications for detoxification. *Int. Rev. Cytol. 260*, 35-112.

Janha, R.E., Worwui, A., Linton, K.J., Shaheen, S.O., Sisay-Joof, F. et al. (2014). Inactive alleles of cytochrome P450 2C19 may be positively selected in human evolution. *BMC Evol. Biol. 14*, 71.

Jiang, Y., Matsunami, H. (2015). Mammalian odorant receptors: functional evolution and variation. *Curr. Opin. Neurobiol. 34*, 54-60.

Kaitin, K.I. (2010). Deconstructing the drug development process: the new face of innovation. *Clin. Pharma. Therap. 87*, 356-361.

Kalgutkar, A.S. (2015). Should the incorporation of structural alerts be restricted in drug design? An analysis of structure-toxicity trends with aniline-based drugs. *Curr. Med. Chem. 22*, 438-464.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M. et al. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic. Acids Res. 40*, D109-D114.

Karatzoglou, A., Smola, A., Hornik, A., Zeileis, A. (2004). kernlab - An S4 package for kernel methods in R. *J. Stat. Softw. 11*, 1-20.

Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I. et al. (2009). Predicting new molecular targets for known drugs. *Nature 462*, 175-181.

Khan, S.A., Faisal, A., Mpindi, J.P., Parkkinen, J.A., Kalliokoski, T. et al. (2012). Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. *BMC Bioinformatics 13*, 112.

Kilpivaara, O., Aaltonen, L.A. (2013). Diagnostic cancer genome sequencing and the contribution of germline variants. *Science 339*, 1559-1562.

Kim, M.T., Huang, R., Sedykh, R., Wang, W., Xia, M. et al. (2016). Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect. 124*, 634-641.

Kirkland, D., Fowler, P. (2010). Further analysis of Ames-negative rodent carcinogens that are only genotoxic in mammalian cells *in vitro* at concentrations exceeding 1mM, including retesting of compounds of concern. *Mutagenesis 25*, 539-553.

Kitagawara, Y, Ohe, T., Tachibana, K., Takahashi, K., Nakamura, S. et al. (2015).Novel bioactivation pathway of benzbromarone mediated by cytochrome P450. Drug Metab. Dispos. 43, 1303-1306.

Kleinstreuer, N. C., Yang, J., Berg, E.L., Knudsen, T.B., Richard, A.M. et al. (2014). Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nat. Biotech. 32*, 583-591.

Klopman, G. (1984). Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc. 106*, 7315-7321.

Kiyosawa, N., Uehara, T., Gao, W., Omura, K., Hirode, M. et al. (2007). Identification of glutathione depletion-responsive genes using phorone-treated rat liver. *J. Toxicol. Sci. 32*, 469-486.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S. et al. (2011). DrugBank 3.0: a comprehensive resource for “omics” research on drugs. *Nucleic. Acids Res. 39*, D1035-D1041.

Kondo, C., Minowa, Y., Uehara, T., Okuno, Y., Nakatsu, N. et al. (2009). Identification of genomic biomarkers for concurrent diagnosis of drug-induced renal tubular injury using a large-scale toxicogenomics database. *Toxicology 265*, 15-26.

Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T.H., von Mering, C. et al. (2014). STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res. 42(Database issue)*, D401-D407.

Kuz'min, V.E., Artemenko, A.G., Muratov, E.N. (2008). Hierarchical QSAR technology based on the Simplex representation of molecular structure. *J. Comput. Aided. Mol. Des. 22*, 403-421.

Kwok, R.H. (1968). Chinese-restaurant syndrome. *N. Engl. J. Med. 278*, 796.

Kwon, H., Park, J., An, Y., Sim, J., Park, S.A. (2014). Smartphone metabolomics platform and its application to the assessment of cisplatin-induced kidney toxicity. *Anal. Chim. Acta. 845*, 15-22.

Lamb, J., Crawford, E.D., Peck, D. Modell, J.W., Blat, I.C. et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science 313*, 1929-1935.

Lang, C., Meier, Y., Stieger, B., Beuers, U., Lang, T. et al. (2007). Mutations and polymorphisms in the bile salt export pump and the multidrug resistance protein 3 associated with drug-induced liver injury. *Pharmacogenet. Genomics 17*, 47-60.

Langford-Smith, A., Keenan, T.D., Clark, S.J., Bishop, P.N., Day, A. J. (2014). The role of complement in age-related macular degeneration: heparin sulphate, a ZIP code for complement factor H? *J. Innate Immun. 6*, 407-416.

Lau, S.S., Zannoni, V.G. (1979). Hepatic microsomal epoxidation of bromobenzene to phenols and its toxicological implication. *Toxicol. Appl. Pharmacol. 50*, 309-318.

Lauschke, V.M., Ingelman-Sundberg, M. (2016). Requirements for comprehensive pharmacogenetic genotyping platforms. *Pharmacogenomics 17*, 917-924.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E. et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature 536*, 285-291.

Lepailleur, A., Poezevara, G., Bureau, R. (2013). Automated detection of structural alerts (chemical fragments) in (eco)toxicology. *Comput. Struct. Biotechnol. J. 5*, e201302013.

Letunic, I., Bork, P. (2011). Interactive Tree of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res. 39*, W475-W478.

Lewell, X.Q., Judd, D.B., Watson, S.P., Hann, M.M. (1998). RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci. 38*, 511–522.

Lindon, J.C., Keun, H.C., Ebbels, T.M., Pearce, J.M., Holmes, E. et al., (2005). The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics 6*, 691-699.

Linge, A., Kennedy, S., O’Flynn, D., Beatty, S., Moriarty, P. et al. (2012). Differential expression of fourteen proteins between uveal melanoma from patients who subsequently developed distant metastases versus those who did not. *Invest. Ophthalmol. Vis. Sci. 53*, 4634-4643.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev. 46*, 3-26

Liu, R., Yu X., Wallqvist, A. (2015). Data-driven identification of structural alerts for mitigating the risk of drug-induced human liver injuries. *J. Cheminform. 7*, 4.

Liu, Y., Jolly, S., Pokala, K. (2013). Prolonged paroxysmal sympathetic storming associated with spontaneous subarachnoid hemorrhage. *Case Rep. Med. 2013*, 358182.

Liu, Z., Fang, H., Borlak, J., Roberts, R., Tong, W. (2017). *In vitro* to *in vivo* extrapolation for drug-induced liver injury using a pair ranking method. *ALTEX 34*, 399-407.

Liu, Z.C., Uetrecht, J.P. (2000). Metabolism of ticlopidine by activated neutrophils: implications for ticlopidine-induced agranulocytosis. *Drug Metab. Dispos. 28*, 726-730.

Lobo, J.M., Jiménez-Valverde, A., Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr. 17*, 145-151.

López-Sagaseta, J., Kung, J.E., Savage, P.B., Gumperz, J., Adams, E.J. (2012). The molecular basis for recognition of CD1d/α-galactosylceramide by a human non-Vα24 T cell receptor. *PLoS Biol. 10*, e1001412.

Lovecka, P., Thimova, M., Grznarova, P., Lipov, J., Knejzlik, Z. et al. (2015). Study of cytotoxic effects of benzonitrile pesticides. *BioMed. Res. Int. 2015*, 381264.

Low Y., Uehara T., Minowa Y., Yamada, H., Ohno, Y. et al. (2011). Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem. Res. Toxicol. 24*, 1251-1262.

Lunde, K., Egelandsdal, B., Skuterud, E., Mainland, J.D., Lea, T. et al. (2012). Genetic variation of an odorant receptor OR7D4 and sensory perception of cooked meat containing androstenone. *PLoS One. 7*, e35259.

Ma, Q., Lu, A.Y.H. (2011). Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacol. Rev. 63*, 437-459.

Mahadevan, B., Snyder, R.D., Waters, M.D., Benz, R.D., Kemper, R.A. et al. (2011). Genetic toxicology in the 21st century: reflections and future directions. *Environ. Mol. Mutagen, 52*, 339-354.

Mainland, J.D., Keller, A., Li, Y.R., Zhou, T., Trimmer, C. et al. (2014). The missense of smell: functional varability in the human odorant receptor repertoire. *Nat. Neurosci. 17*, 114-120.

Martin, Y.C., Kofron, J.L., Traphagen, L.M. (2002). Do structurally similar molecules have similar biological activity? *J. Med. Chem. 45*, 4350-4358.

Mazzatorta, P., Estevez, M.D., Coulet, M., Schilter, B. (2008). Modeling oral rat chronic toxicity. *J. Chem. Inf. Model 48*, 1949-1954.

McDonald, M.G., Rettie, A.E. (2007). Sequential metabolism and bioactivation of the hepatotoxin benzbromarone: formation of glutathione adducts from a catechol intermediate. *Chem. Res. Toxicol. 20*, 1833-1842.

McGregor, D., Lang, M. (2000). Carbon tetrachloride: genetic effects and other modes of action. *Mutat. Res. 366*, 181-195.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R. et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biol. 17*, 122.

McLeod, H.L., Evans, W.E. (2001). Pharmacogenomics: unlocking the human genome for better drug therapy. *Annu. Rev. Pharmacol. Toxicol. 41*, 101-121.

Mellora, C.L., Schultzb, T.W., Przybylaka, K.R., Richarza, A.N., Bradbury, S.P., et al. (2017). Read-across for rat oral gavage repeated-dose toxicity for short-chain mono-alkylphenols: A case study. *Comput. Toxicol. 2*, 1-11.

Merico, D., Isserlin, R., Stueker, O., Emili, A., Bader, G.D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One 5*, e13984.

Mésange, F., Sebbar, M., Capdevielle, J., Guillemot, J.C., Ferrara, P. et al. (2002). Identification of two tamoxifen target proteins by photolabeling with 4-(2-morpholinoethoxy)benzophenone. *Bioconjug. Chem. 13*, 766-772.

Montmayeur, J.P., Matsunami, H. (2002). Receptors for bitter and sweet taste. *Curr. Opin. Neurobiol. 12*, 366-371.

Moro, S., Chipman, J.K., Antczak, P., Turan, N., Dekant, W. et al. (2012). Identification and pathway mapping of furan target proteins reveal mitochondrial energy production and redox regulation as critical targets of furan toxicity. *Toxicol. Sci. 126*, 336-352.

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Müller, L., Pähler, A. (2008). Computational toxicology in drug development. *Drug Discov. Today 13*, 303–310.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., Prachayasittikul, V. (2010). Advances in computational methods to predict the biological activity of compounds. *Expert. Opin. Drug Discov. 5*, 633-654.

Nebert, D.W. (1997). Polymorphisms in drug-metabolizing enzymes: what is their clinical relevance and why do they exist? *Am. J. Hum. Genet. 60*, 265-271.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P. et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science 337*, 100-104.

Nelson, S.D. (2001). Structure toxicity relationships--how useful are they in predicting toxicities of new drugs? *Adv. Exp. Med. Biol. 500*, 33-43.

Niimura, Y., Nei, M. (2006). Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *J. Hum. Genet. 51*, 505-517.

O'Boyle, N.M., Banck, M., James, C.A., Morley C, Vandermeersch, T. et al. (2011). Open Babel: An open chemical toolbox. *J. Cheminform. 3*, 33.

Omura, K., Kiyosawa, N., Uehara, T., Hirode, M., Shimizu, T. et al. (2007). Gene expression profiling of rat liver treated with serum triglyceride-decreasing compounds. *J. Toxicol. Sci. 32*, 387-399.

Onakpoya, I.J., Heneghan, C.J., Aronson, J.K. (2016). Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. *BMC Med. 14*, 10.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L. et al. (2014). The MIntAct project -- IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic. Acids Res. 42(Database issue)*, D358-D363.

Paul, S.M., Mytelka, D.S., Eunwiddie, C.T., Persinger, C.C., Munos, B.H. et al. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Rev. Drug Discov. 9*, 203-214.

Pauwels, E., Stoven, V., Yamanishi, Y. (2011). Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics 12*, 169.

Perualila-Tan, N., Kasim, A., Talloen, W., Verbist, B., Göhlmann, H.W. et al. (2016). A joint modeling approach for uncovering associations between gene expression, bioactivity and chemical structure in early drug discovery to guide lead selection and genomic biomarker development. Stat. *Appl. Genet. Mol. Biol. 15*, 291-304.

Pires, D.E., Ascher, D.B., Blundell, T.L. (2014). DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic. Acids Res. 42*, W314-W319.

Pires, D.E., Blundell, T.L., Ascher, D.B. (2015). Platinum: a database of experimentally measured effects of mutations on structurally defined protein-ligand complexes. *Nucleic. Acids Res. 43*, D387-D391.

Pires, D.E., Blundell, T.L., Ascher, D.B. (2016). mCSM-lig: quantifying the effects of mutations on protein-small molecule affinity in genetic disease and emergence of drug resistance. *Sci. Rep. 6*, 29575.

Ploemen, J.H., van Ommen, B., Bogaards, J.J., van Bladeren, P.J. (1993). Ethacrynic acid and its glutathione conjugate as inhibitors of glutathione S-transferases. *Xenobiotica 23*, 913-923.

Pronin, A.N., Xu, H., Tang, H., Zhang, L., Li, Q., Li, X. (2007). Specific alleles of bitter receptor genes influence human sensitivity to the bitterness of aloin and saccharin. *Curr. Biol. 17*, 1403-1408.

Puga, A., Nebert, D., McKinnon, R., Menon, A. (1997). Genetic polymorphisms in human drug-metabolizing enzymes: potential uses of reverse genetics to identify genes of toxicological relevance. *Crit. Rev. Toxicol. 27*, 199-222.

Poroikov, V.V., Filimonov, D.A., Borodina, Y.V., Lagunin, A.A., Kos, A. (2000). Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci. 40*, 1349-1355.

R Core Development Team (2012). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*, http://www.R-project.org/

Rebsamen, M.C., Desmeules, J., Daali, Y., Chiappe, A., Diemand, A. et al. (2009). The AmpliChip CYP450 test: cytochrome P450 2D6 genotype assessment and phenotype prediction. *Pharmacogenomics J. 9*, 34-41.

Relling, M.V., Gardner, E.E., Sandborn, W.J., Schmiegelow, K., Pui, C.H. et al. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. *Clin. Pharmacol. Ther. 89*, 387-391.

Reva, B., Antipin, Y., Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic. Acids Res. 39*, e118.

Ridings, J.E., Barratt, M.D., Cary, R., Earnshaw, C.G., Eggington, C.E. et al. (1996). Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology 106*, 267-279.

Roses, A.D. (2000). Pharmacogenetics and the practice of medicine. *Nature 405*, 857-865.

Roses, A.D. (2002). Genome-based pharmacogenetics and the pharmaceutical industry. *Nature Rev. Drug Discov. 1*, 541-549.

Roth, B., Lopez, E., Patel, S., Kroeze, W. (2000). The multiplicity of serotonin receptors:

uselessly diverse molecules or an embarrassment of riches? *Neuroscientist*. *6*,262.

Roth, B.L. (2007). Drugs and vascular heart disease. *N. Engl. J. Med. 356*, 6-9.

Rothman, R.B., Baumann, M.H., Savage, J.E., Rauser, L., McBride, A. et al. (2000). Evidence for possible involvement of 5-HT(2B) receptors in the cardiac valvulopathy associated with fenfluramine and other serotonergic medications. *Circulation 102*, 2836-2841.

Roy, B., Chowdhury, A., Kundu, S., Santra, A., Dey, B. et al. (2001). Increased risk of antituberculosis drug-induced hepatotoxicity in individuals with glutathione S-transferase M1 'null' mutation. *J. Gastroenterol. Hepatol. 16*, 1033-1037.

Rybacka, A., Rudén, C., Tetko, I. V., Andersson, P. L. (2015). Identifying potential endocrine disruptors among industrial chemicals and their metabolites-development and evaluation of *in silico* tools. *Chemosphere 139*, 372-378.

Rydberg, P., Gloriam, D.E., Zaretzki, J., Breneman, C, Olsen, L. (2010). SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett. 1*, 96-100.

Schnellmann, R.G., Mandel, L.J. (1986). Cellular toxicity of bromobenzene and bromobenzene metabolites to rabbit proximal tubules: the role and mechanism of 2-bromohydroquinone. *J. Pharmacol. Exp. Ther. 237*, 456-461.

Schnur, D.M., Hermsmeier, M.A., Tebben, A.J. (2006). Are target-family-privileged substructures truly privileged? *J. Med. Chem. 49*, 2000-2009.

Schoch, G.A., Yano, J.K., Sansen, S., Dansette, P.M., Stout, C.D. et al. (2008). Determinants of cytochrome P450 2C8 substrate binding: structures of complexes with montelukast, troglitazone, felodipine, and 9-cis-retinoic acid. *J. Biol. Chem. 283*, 17227-17237.

Schuffenhauer, A., Ruedisser, S., Marzinzik, A.L., Jahnke, W., Blommers, M. et al. (2005). Library design for fragment based screening. *Curr. Top. Med. Chem. 5*, 751-762.

Sedykh, A.Y., Klopman, G. A. (2006). Structural analogue approach to the prediction of the octanol-water partition coefficient. *J. Chem. Inf. Model. 46*, 1598-1603.

Sevrioukova, I.F., Poulos, T.L. (2012). Structural and mechanistic insights into the interaction of cytochrome P4503A4 with bromoergocryptine, a type I ligand. *J. Biol. Chem. 287*, 3510-3517.

Shah, A., Shinde, R., Kare, P., Hymavathi, V., Chavan, S. et al. (2012). Induced fit binding of aldose reductase inhibitors to AKR1B10. *Med. Chem. Res. 21*, 1245-1252.

Sheng, X., Zhang, L., Tong, N., Luo, D., Wang, M. et al. (2012). MDR1 C3435T polymorphism and cancer risk: a meta-analysis based on 39 case-control studies. *Mol. Biol. Rep. 39*, 7237-7249.

Shi, P., Zhang, J., Yang, H., Zhang, Y.P. (2003). Adaptive diversification of bitter taste receptor genes in Mammalian evolution. *Mol. Biol. Evol. 20*, 805-814.

Shruthi, B.S., Vinodhkumar, P., Selvamani. (2016). Proteomics: a new perspective for cancer. *Adv. Biomed. Res. 5*, 67.

Siegel, M.G., Vieth, M. (2007). Drugs in other drugs: a new look at drugs as fragments.

Drug Discov. Today 12, 71-79.

Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. et al. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic. Acids Res. 40(Web Server issue)*, W452-W457.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics 21*, 3940-3941.

Singh, A.K., Kumar, R.P., Pandey, N., Singh, N., Sinha, M. et al. (2010). Mode of Binding of the Tuberculosis Prodrug Isoniazid to Heme Peroxidases. *J. Biol. Chem. 285*, 1569-1576.

Siramshetty, V.B., Nickel, J., Omieczynski, C., Gohlke, B., Drwal, M.N. et al. (2016). WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic. Acids Res. 44(D1)*, D1080-D1086.

Somberg, J.C., Molnar, J. (2009). The pleiotropic effects of ethacrynic acid. *Am. J. Ther. 16*, 102-104.

Speir, J.A., Abdel-Motal, U., Jondal, M., Wilson, I.A. (1999). Crystal structure of an MHC class I presented glycopeptide that generates carbohydrate-specific CTL. *Immunity 10*, 51-61.

Spyridopoulou, K.P., Dimou, N.L., Hamodrakas, S.J., Bagos, P.G. (2012). Methylene tetrahydrofolate reductase gene polymorphisms and their association with methotrexate toxicity: a meta-analysis. *Pharmacogenet. Genomics 22*, 117-133.

Stepan, A.F., Walker, D.P., Bauman, J., Price, D.A., Baillie, T.A. et al. (2011). Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol. 24*, 1345-1410.

Storey, J.D., Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A. 100*, 9440-9445.

Sun, L., Song, L., Wan, Q., Wu, G., Li, X., et al. (2015). cMyc-mediated activation of serine biosynthesis pathway is critical for cancer progression under nutrient deprivation conditions. *Cell Res. 25*, 429-444.

Takahashi, H., Echizen, H. (2003). Pharmacogenetics of CYP2C9 and interindividual variability in anticoagulant response to warfarin. *Pharmacogenomics J. 3*, 202-214.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N. et al. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet. 5*, e1000433.

Tanaka, K., Kiyosawa, N., Watanabe, K., Manabe, S. (2007). Characterization of resistance to bromobenzene-induced hepatotoxicity by microarray. *J. Toxicol. Sci. 32*, 129-134.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature 526*, 68-74.

Thompson, D.C., Perera, K., London, R. (1995). Quinone methide formation from para isomers of methylphenol (cresol), ethylphenol, and isopropylphenol: relationship to toxicity. *Chem. Res. Toxicol. 8*, 55-60.

Thor, H., Orrenius, S. (1980). The mechanism of bromobenzene-induced cytotoxicity studied with isolated hepatocytes. *Arch. Toxicol. 44*, 31-43.

Todeschini, R., Consonni, V. (2000). *Handbook of Molecular Descriptors, Wiley-VCH.*

Trowsdale, J., Parham, P. (2004). Mini-review: defense strategies and immunity-related genes. *Eur. J. Immunol. 34*, 7-17.

Uehara, T., Hirode, M., Ono, A., Kiyosawa, N., Omura, K. et al. (2008a). A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats. *Toxicology 250*, 15-26.

Uehara, T., Kiyosawa, N., Hirode, M., Omura, K., Shimizu, T. et al. (2008b). Gene expression profiling of methapyrilene-induced hepatotoxicity in rat. *J. Toxicol. Sci. 33*, 37-50.

Uehara, T., Minowa, Y., Morikawa, Y., Kondo, C., Maruyama, T. et al. (2011). Prediction model of potential hepatocarcinogenicity of rat hepatocarcinogens using a large-scale toxicogenomics database. *Toxicol. Appl. Pharmacol. 255*, 297-306.

Uehara, T., Ono, A., Maruyama, T., Kato, I., Yamada, H. et al. (2010). The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res. 54*, 218-227.

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic. Acids Res. 43(Database issue)*, D204-D212.

Unterthiner, T., Mayr, A., Klambauer, G., Hochreiter, S. (2016). Toxicity prediction using deep learning. *Available online*: http;//arXiv Prepr arXiv1503.01445.

Valerio, L.G. Jr. (2009). *In silico* toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol. 241*, 356-370.

Vedani, A., Dobler, M., Smiesko, M. (2012). VirtualToxLab - a platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol. Appl. Pharmacol. 261*, 142-153.

Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J. et al. (1999). Analysis of human transcriptomes. *Nat. Genet. 23*, 387-388.

Walker, B.D., Singleton, C.B., Bursill, J.A., Wyse, K.R., Valenzuela, S.M. et al. (1999). Inhibition of the human ether-a-go-go-related gene (HERG) potassium channel by cisapride: affinity for open and inactivated states. *Br. J. Pharmacol. 128*, 444-450.

Waring, J.F., Jolly, R.A., Ciurlionis, R., Lum, P.Y., Praestgaard, J.T. et al. (2001). Clustering of hepatotoxins based on mechanism of toxicity using gene expression profiles. *Toxicol. Appl. Pharmacol. 175*, 28-42.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci. 28*, 31-36.

Williams, D.P., Park, B.K. (2003). Idiosyncratic toxicity: the role of toxicophores and bioactivation. *Drug Discov. Today 8*, 1044-1050.

Wilde, S., Timpson, A., Kirsanow, K., Kaiser, E., Kayser, M. (2014). Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl. Acad. Sci. U S A. 111*, 4832-4837.

Wilson, C.L., Miller, C.J. (2005). Simpleaffy: a BioConductor package for Affymetrix

Quality Control and data analysis. *Bioinformatics 21*, 3683-3685.

Witten, D.M., Tibshirani, R., Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*, 515-534.

Wong, A., Soo, R.A., Yong, W.P., Innocenti, F. (2009). Clinical pharmacology and pharmacogenetics of gemcitabine. *Drug Metab. Rev. 41*, 77-88.

Xenarios, I., Salwínski, L., Duan, X.J., Higney, P., Kim, S.M. et al. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic. Acids Res. 30*, 303-305.

Yamada, F., Sumida, K., Uehara, T., Morikawa, Y., Yamada, H. et al. (2012). Toxicogenomics discrimination of potential hepatocarcinogenicity of non-genotoxic compounds in rat liver. *J. Appl. Toxicol. 33*, 1284-1293.

Yamane, J., Aburatani, A., Imanishi, S., Akanuma, H., Nagano, R. et al. (2016). Prediction of developmental chemical toxicity based on gene networks of human embryonic stem cells. *Nucleic. Acids Res. 44*, 5515-5528.

Yamtich, J., Speed, W.C., Straka, E., Kidd, J.R., Sweasy, J.B. et al. (2009). Population-specific variation in haplotype composition and heterozygosity at the POLB locus. *DNA Repair 8*, 579-584.

Zhang, C., Cheng, F., Li, W., Liu, G., Lee, P.W. et al (2016). *In silico* prediction of drug induced liver toxicity using substructure pattern recognition method. *Mol. Informatics 35*, 136-144.

Zhang, M., Wang, Y., Wang, Q., Yang, J., Yang, D. et al. (2010). Involvement of mitochondria-mediated apoptosis in ethylbenzene-induced renal toxicity in rat. *Toxicol. Sci. 115*, 295-303.

Zhang, S., Liu, S., Tao, R., Wei, D., Chen, L. et al. (2012). A highly selective and potent PTP-MEG2 inhibitor with therapeutic potential for type 2 diabetes. *J. Am. Chem. Soc. 134*, 18116-18124.

Zhong, S.L., Zhou, S.F., Chen, X., Chan, S.Y., Chan, E. et al. (2006). Relationship between genotype and enzyme activity of glutathione S-transferases M1 and P1 in Chinese. *Euro. J. Pharm. Sci. 28*, 77-85.

Zhu, X., Dickerson, T.J., Rogers, C.J., Kaufmann, G.F., Mee, J.M. et al. (2006). Complete reaction cycle of a cocaine catalytic antibody at atomic resolution. *Structure 14*, 205-216.

Supporting Information

The supplementary tables and files can be found at: http://www.russelllab.org/YYJ_thesis/

Chapter II:

Table S2.1 Toxicity structural alerts and their frequency of occurrence in each of the chemical classes

Table S2.2 Performance for the training and external validation set using SVM

Table S2.3 Extracted chemical fragments and toxicities by SCCA

Table S2.4 Prediction results for withdrawn drugs by SCCA

Chapter III:

Table S3.1 Comparison of datasets used in this study

Table S3.2 The 100 strongest deregulated genes (229 probe sets) at the highest concentration for the incubation period of 24h *in vitro* rat hepatocyte across all compounds

Table S3.3 Pathology class of 131 compounds used for treatment of rat liver and rat kidney

Table S3.4 100 strongest dysregulated genes across all chemicals *in vitro* rat hepatocytes

Table S3.5 The expression levels of the individual genes upregulated at three concentrations after 24h exposure *in vitro* rat hepatocytes

Table S3.6 Gene regulated among ligands of peroxisome proliferator-activated receptor in *in vivo* rat liver and *in vitro* rat and human hepatocytes

Table S3.7 Gene regulated among musculoskeletal drugs in *in vivo* rat liver and *in vitro* rat and human hepatocytes

Table S3.8 Gene regulated among drugs acting on the nervous system in in vivo rat liver and *in vitro* rat and human hepatocytes

Table S3.9 Gene regulated among hepatotoxicants in *in vivo* rat liver

Table S3.10 Gene regulated among antineoplastic and immunomodulating agents in *in vivo* rat liver and *in vitro* rat and human hepatocytes

Table S3.11 Gene regulated among drugs acting on the alimentary tract and metabolism in *in vivo* rat liver and *in vitro* rat and human hepatocytes

Table S3.12 Gene regulated among chemicals *in vivo* rat liver and *in vivo* rat kidney

Table S3.13 Chemical descriptors of chemicals that dysregulated gene expression

Table S3.14 Correlations between gene expression similarity and molecular structure similarity

Table S3.15 Associations of chemical fragments and *in vivo* rat liver pathological outcomes extracted by SCCA

Table S3.16 Associations of gene expressions and pathological observations from *in vivo* rat liver experiment extracted by SCCA

Table S3.17 Associations of gene expressions and pathological observations from *in vivo* rat kidney experiment extracted by SCCA

Table S3.18 Associations of gene expressions and pathological observations from *in vitro* rat hepatocyte experiment extracted by SCCA