TOXICITY PREDICTION
USING MULTI-DISCIPLINARY DATA INTEGRATION
AND NOVEL COMPUTATIONAL APPROACHES

Yen Sia Low

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the Department of Environmental Sciences and Engineering,
Gillings School of Global Public Health.

Chapel Hill
2013

Approved by:

Alexander Tropsha, Ph.D.

Ivan Rusyn, M.D., Ph.D.

Louise Ball, Ph.D.

Avram Gold, Ph.D.

David Dix, Ph.D.

David Leith, Ph.D.

ABSTRACT

YEN SIA LOW: TOXICITY PREDICTION USING MULTI-DISCIPLINARY DATA INTEGRATION AND NOVEL COMPUTATIONAL APPROACHES
(Under the direction of Alexander Tropsha and Ivan Rusyn)

Current predictive tools used for human health assessment of potential chemical hazards rely primarily on either chemical structural information (i.e., cheminformatics) or bioassay data (i.e., bioinformatics). Emerging data sources such as chemical libraries, high throughput assays and health databases offer new possibilities for evaluating chemical toxicity as an integrated system and overcome the limited predictivity of current fragmented efforts; yet, few studies have combined the new data streams.

This dissertation tested the hypothesis that integrative computational toxicology approaches drawing upon diverse data sources would improve the prediction and interpretation of chemically induced diseases. First, chemical structures and toxicogenomics data were used to predict hepatotoxicity. Compared with conventional cheminformatics or toxicogenomics models, interpretation was enriched by the chemical and biological insights even though prediction accuracy did not improve. This motivated the second project that developed a novel integrative method, chemical-biological read-across (CBRA), that led to predictive and interpretable models amenable to visualization. CBRA was consistently among the most accurate models on four chemical-biological data sets. It highlighted chemical and biological features for interpretation and the visualizations aided transparency.

Third, we developed an integrative workflow that interfaced cheminformatics prediction with pharmacoepidemiology validation using a case study of Stevens Johnson Syndrome (SJS), an adverse drug reaction (ADR) of major public health concern. Cheminformatics models first predicted potential SJS inducers and non-inducers, prioritizing them for subsequent pharmacoepidemiology evaluation, which then confirmed that predicted non-inducers were statistically associated with fewer SJS occurrences. By combining cheminformatics' ability to predict SJS as soon as drug structures are known, and pharmacoepidemiology's statistical rigor, we have provided a universal scheme for more effective study of SJS and other ADRs.

Overall, this work demonstrated that integrative approaches could deliver more predictive and interpretable models. These models can then reliably prioritize high risk chemicals for further testing, allowing optimization of testing resources. A broader implication of this research is the growing role we envision for integrative methods that will take advantage of the various emerging data sources.

This dissertation is dedicated to my family, especially my parents,

whose unwavering love and support throughout my life

have encouraged me to dream and made this research possible.

ACKNOWLEDGEMENTS

I wish to thank the following people who have enriched my life as a doctoral student at UNC-Chapel Hill. First and foremost, I thank advisors Alex Tropsha and Ivan Rusyn whose generous support have allowed me to pursue interests beyond cheminformatics. More importantly, their gift of time rounded my scientific apprenticeship during which I was invited to observe, participate and contribute to the discourse between two great minds.

I thank members of the Molecular Modeling Lab (MML) whose presence have made MML a wonderful place to work and play. I am grateful for their friendship and generous feedback without which my research experience would have been much poorer. In particular, I thank Alec Sedykh, Denis Fourches, Sasha Golbraikh and Regina Politi for being such patient sounding boards to my ideas both good and bad. I also thank dear friends Liying Zhang and Guiyu Zhao for their company on too many late nights.

I am grateful to friends near and afar for their inspiration and support, filling my PhD journey with great memories. I thank my family for always believing in me. Thank you Mummy for being the voice of reason and comfort. A special thank you too, Daniel Oreper, my boyfriend, for patiently sharing in my joys and pains. Finally, I thank the grace of God for putting all the above people in my life to make this dissertation possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Applicability domain
ADME	Absorption, distribution, metabolism, and excretion
ADR	Adverse drug reaction
ALP	Alkaline phosphatase
ALT	Alanine transaminase
AST	Aspartate transaminase
ATC	Anatomical therapeutic chemical
AUC	Area under curve
CAS	Chemical Abstracts Service
CBRA	Chemical-biological read-across
CCA	Canonical correlation analysis
CCR	Correct classification rate (also termed balanced accuracy)
CPDB	Carcinogenicity potency database
CV	Cross-validation
CWAS	Chemistry-wide association studies
DBIL	Direct bilirubin
DNA	Deoxyribonucleic acid
DWD	Distance-weighted discrimination
ER	Endoplasmic reticulum
FDA	Food and Drug Administration
GGT	Gamma-glutamyl transpeptidase
GWAS	Genome-wide association studies
HTS	High throughput screening

ISIDA	In Silico Design and Data Analysis
*k*NN	*k* nearest neighbors
MACCS	Molecular ACCess System
MCS	Maximal common substructure
MeSH	Medical subject heading
MOE	Molecular Operating Environment
OECD	Organisation for Economic Co-operation and Development
OR	Odds ratio
PCA	Principal component analysis
PPAR	Peroxisome proliferator activated receptor
(Q)SAR	(Quantitative) structure-activity relationship
RA	Read-across
REACH	Registration, Evaluation and Authorization of CHemicals
RF	Random forests
SA	Structural alert
SAM	Significance analysis of microarrays
SD	Standard deviation
SiRMS	Simplex Representation of Molecular Structure
SJS	Stevens Johnson syndrome
SOM	Self-organizing map
SVM	Support vector machines
TBIL	Total bilirubin
WHO	World Health Organization

LIST OF SYMBOLS

A_i	Observed activity of *i*th neighbor
A_pred	Predicted activity of compound
d_Jac	Jaccard distance between 2 compounds
$I(x,\mathrm{compound})$	Local importance score of descriptor *x* with respect to compound
k	Maximum number of neighbors in *k* nearest neighbors
k_bio	Maximum number of biological neighbors in *k* nearest neighbors
k_chem	Maximum number of chemical neighbors in *k* nearest neighbors
r^2	Pearson correlation coefficient between 2 descriptors
S_i	Similarity between a compound and its *i*th neighbor
x_m	*m*th descriptor value of a compound

CHAPTER 1. INTRODUCTION

Contrary to popular belief, most of the 83,000 chemicals in use in the US under the Toxic Substances Control Act have not been extensively tested. Adding to this backlog are another 700 chemicals introduced each year (Stephenson 2009). Consequently, there is an urgent need for smarter toxicity testing strategies to facilitate timely and informed decisions. Central to this is a tiered approach which funnels the chemicals through *in silico, in vitro* and *in vivo* tests in order of decreasing throughput (Dix et al. 2007, Keller et al. 2012, Merlot 2008).

Among the *in silico* methods, cheminformatics and bioinformatics have established themselves as integral parts of toxicity testing, especially in the initial stages where their high throughput advantage comes into play. The following sections will describe their current roles in toxicity testing and lay down a framework for an integrative chemical-biological approach that I posit will improve toxicity assessment in terms of higher accuracy and richer interpretation.

1.1. Chemical structural data and cheminformatics in predictive toxicology

Because cheminformatics models require only chemical structures as inputs for modeling, they allow prediction as soon as the molecule is designed *in silico* and are increasingly seen as the *de facto* tools during the first stages of toxicity testing. Among the best known cheminformatics tools in widespread use are OECD Toolbox, Derek expert system, and MultiCASE, drawing upon various cheminformatics methods detailed below (Gatnik & Worth 2010).

Cheminformatics methods may be broadly classified into ligand-based or structure-based approaches depending on the reference frame used to define the molecule of interest (small molecule *ligand* or large biomolecule *structure*). Structure-based approaches are more commonly used in drug discovery where the drug target structure (e.g., a protein receptor) is defined. Ligand-based approaches are preferred in toxicity assessment especially when the exact molecular targets mediating toxicity may not have been elucidated.

Cheminformatics-based toxicity prediction relies heavily on ligand-based approaches, especially quantitative structure-activity relationship (QSAR) modeling. QSAR models which relate small molecule structure to chemical activity through statistical functions first appeared in 1962 when Hansch et al. correlated growth activity of auxins with their molecular electronic properties (Hansch et al. 1962). It built upon the research in descriptor development quantifying key molecular properties, most notably, electronic (Hammett 1937), hydrophobic (Collander et al. 1951) and steric (Taft 1952) properties (Selassie & Verma 2010). Since then, the number of chemical descriptors has grown into the thousands in our attempt to comprehensively characterize molecules. For a thorough discussion of descriptors, refer to (Selassie & Verma 2010, Todeschini & Consonni 2000). The diverse range of descriptors covers physicochemical properties, substructural fragments (*e.g.* presence of chemical function groups), molecular signatures (*e.g.* MACCS fingerprints) and abstract mathematical derivations based on quantum theory (*e.g.* orbital energies). Parallel to the development of descriptors, modeling methods have evolved from simple linear regression to complex machine learning algorithms.

Arising from this multitude of descriptors and methods are (Q)SAR and related variants to suit various user needs (Gleeson et al. 2012). The simplest and most interpretable among them is the structural alert or toxicophore whose presence in a molecule acts as a heuristic indicator for toxicity. Structural alerts may be expert-derived (Ashby 1978) or empirically mined (Rosenkranz & Klopman 1988). The Derek expert system uses a collection of structural alerts to predict mutagenicity. The next simplest is SAR which, unlike QSAR, *qualitatively* relates chemical features to toxicity. Another widely used technique is read-across which infers toxicity from the known toxicity outcomes of similar chemicals. Here, chemical similarity may be defined qualitatively (e.g. presence of substructures) or quantitatively (e.g. Tanimoto similarity coefficients calculated from chemical descriptors). The OECD Toolbox is powered by a mix of (Q)SAR and read-across methods.

Generally, quantitative modeling with more descriptors increases predictivity but requires the use of more complicated modeling methods which obscure interpretability. Hence, the choice of tool for toxicity estimation depends on the user’s needs. A medicinal chemist may opt for a high accuracy QSAR model to eliminate unpromising drug candidates while a regulator may prefer an interpretable read-across method whose transparency fulfills documentation requirements and sheds light on toxicological mechanisms. However, pitting predictivity against interpretability presents a false dichotomy as the two are inextricably inter-dependent: interpretation is conditional on predictivity; prediction without explanatory power raises doubts on its scientific validity (Shmueli 2010).

One may argue that the common criticism of QSAR models being black boxes is unjustified. Perhaps, the limitation is the unfortunate consequence of QSAR’s success as its ease of use has encouraged blind application among practitioners who lack the fundamental understanding in chemistry and statistics for model interpretation. In defense of QSAR modeling, Chapter 4 will show how the appropriate choice of chemical descriptors and modeling methods can achieve both predictivity and interpretability.

Nevertheless, limitations in data quality, training set selection, modeling methods, validation procedures and model interpretation pose challenges. Countermeasures include careful data curation (Fourches et al. 2010), appropriate data treatment (Eriksson et al. 2003), representative sampling of the chemical space (Golbraikh & Tropsha 2002a), high dimensional modeling techniques, stringent validation (Tropsha & Golbraikh 2007) and rationale-driven descriptor selection to simplify interpretation (Scior et al. 2009). Such attempts to standardize and improve (Q)SAR modeling practices have led to the development of guidelines (OECD 2007) and best practices for QSAR modeling (Tropsha 2010). The five OECD principles for good (Q)SAR modeling are: (1) a well-defined endpoint, (2) unambiguous algorithm, (3) defined applicability domain, (4) appropriate measures of model performance, and (5) mechanistic interpretation where possible.

Despite the above measures, cheminformatics-based prediction of complex toxic phenomena has fallen short of expectation. In reality, the relationship between chemical structures and toxicity is far more circuitous than the models assume, involving many non-chemical factors including those dependent on the biological host (e.g. toxicokinetics, repair capacity). The significance of these non-chemical factors depends on the prediction target. Generally, QSAR models are more successful at predicting direct chemical-induced outcomes (e.g. mutagenicity) than outcomes farther downstream of chemical-initiating events (e.g. carcinogenicity). While QSAR models for mutagenicity (largely molecular interactions between chemical and DNA) approach the accuracy of the Ames assay, carcinogenicity has been notoriously difficult to predict because of its heterogeneous modes of action and the biological host's adaptive capacity for recovery (Benigni 2005). One way to account for these biological factors is to formulate them into the QSAR models.

1.2. Bioassay data and bioinformatics in predictive toxicology

The post-genome era saw a shift towards molecular toxicology and the corresponding rise of bioinformatics. The field of bioinformatics is broad, involving the computational analysis of biological information arising from the detailed characterization of an organism at various levels (molecular, cellular, tissue, organ, system). While bioinformatics has many subdisciplines (e.g. sequencing, 'omics, systems biology), this section focuses on a subset with toxicology applications where the goal is to systematically study multiple biological perturbations in response to chemical insult.

Two broad applications for bioinformatics are toxicity prediction and mechanistic elucidation. Prediction attempts to forecast long-term toxicity endpoints such as cancer from short-term assay surrogates, while elucidation is more concerned with explaining complex toxicological phenomena in terms of simpler biological entities. In addition to forecasting, predicting toxicity from a reduced battery of assays allows researchers to focus their testing resources.

Regardless of the objective, large-scale bioassay data is first required. Fortunately, advances in assay technology has given rise to a diversity of biological measures such as 'omics (e.g. transcriptomics, proteomics, metabonomics), enzymatic activity, receptor binding affinity, cytotoxicity, and histology imaging, allowing toxicologists to probe into both microscopic and macroscopic changes in the body. These bioassays may have different predictive power depending on the experimental error and biological relevance. High-dimensional ‘omics, especially transcriptomics, were shown to have high predictive value (Afshari et al. 2011, Chen et al. 2012, Heijne et al. 2005). Long term toxicity endpoints such as 2-year hepatic tumorigenicity were successfully predicted from short-term 1-, 3-, 5-day transcriptomics (Fielden et al. 2007). The same group also identified 35 gene expression markers important for predicting nephrotoxicity in another study (Fielden et al. 2005), obviating the need for subsequent full microarray analysis. Others such as ToxCast assays, capturing a large diversity of biological characteristics, were less predictive (Thomas et al. 2012).

Simultaneously studying thousands of bioassays offers several advantages: key biomarkers can be quickly identified and interactions between them characterized, allowing a systems toxicology approach. In drug discovery, the use of diverse bioassay panels helps to quickly identify potentially toxic properties (e.g. cytochrome P450 inhibition, transporter blockage) which may be clues into the pathogenesis of a compound. The bioassay signatures of compounds exemplifying certain toxic modes of action may be used to probe for similarly acting compounds. An example is the Japanese Toxicogenomics Project which ascertained toxicogenomic signatures representative of various types of hepatotoxicities (e.g. phospholipidosis, glutathione depletion) for which drugs with unknown hepatoxicities may be measured against (Uehara et al. 2010).

However, bioinformatics is not without criticism. The ease of collecting large-scale bioassay data has encouraged fishing expeditions which tax the limits of current computational methods, leading to false discoveries. Applied to thousands of assays, small probabilities due to chance translate to multiple “discoveries”. Overly sensitive ‘omics markers may be more noise than signal (Zhang et al. 2012). Countermeasures include proper statistical correction (e.g. Bonferroni, Holm) and proper application of biological context to draw meaningful conclusions from the data.

Fortunately, the interpretation of key biological events underlying toxicity is aided by numerous curated databases such as the Comparative Toxicogenomics Database (http://ctdbase.org/), Connectivity Map (http://www.broadinstitute.org/cmap/) and Ingenuity Knowledge Base (http://www.ingenuity.com/) which maps functional genomics markers associated with chemicals and toxicological phenotypes onto functional pathways. As much as these knowledge repositories have made functional analysis more accessible, the generation of new insight still requires a profound understanding of toxicology. Fundamental differences in the way biological processes are organized at the various levels within an organism dictate the extent of *in vitro-in vivo* extrapolation. For instance, the absence of metabolism in *in vitro* systems means that subcellular changes are unlikely to be representative of whole animal phenotypes involving metabolic activation or metabolic clearance (Kienhuis et al. 2009). Therefore, the importance of biological expertise in guiding interpretation cannot be overstressed.

In efforts to improve interpretation, some studies employ complementary technologies. Of note is one multi-omics study which reported that separate genomics, proteomics and metabolomics analyses mutually validate one another's findings and point towards common biological processes consistent with methapyrilene-induced hepatotoxicity (Craig et al. 2006).

The focus on biological information, due to the nature of bioinformatics, has regrettably overlooked another important component of toxicology: chemical information. While bioassays were previously performed for a few chemicals due to throughput limitations, it is now possible to perform HTS on numerous chemicals. Consequently, toxicity data is rich in both biological and chemical information. The underlying chemical patterns, a rich data source for modeling as demonstrated by cheminformatics, have not been capitalized upon by bioinformatics. A reasonable approach may be to complement bioinformatics with cheminformatics for improved toxicity prediction. This leads to the following thesis that an integrative chemical-biological approach will benefit toxicity prediction in terms of predictivity and interpretability.

1.3. Motivation for integrative chemical-biological modeling in predictive toxicology

Given cheminformatics' inadequate consideration of biological factors and bioinformatics' non-use of chemical structures, the concurrent study of both biological and chemical domains may uncover new insights previously invisible to either domain alone. Such integrated approaches attempt to formulate chemical toxicity as a system of interconnected chemical and biological entities. Toxicity, whether occurring at the molecular, cellular, or systemic level, originates from a complex interplay between the chemical inducer and the biological host. Chemical factors govern the molecular interactions between the chemical and its protein targets. The molecular interactions then initiate a cascade of interactions within the cell, organ or organism, eventually giving rise to the observed toxicity phenotype.

Moreover, the rise of several recent enabling trends facilitates chemical-biological integration. First, there is an increased demand for *in silico* and *in vitro* tests instead of *in vivo* tests in efforts to boost testing throughput, improve animal welfare and deepen our understanding of the toxicological mechanisms, accelerated by recent initiatives such as REACH (Registration, Evaluation and Authorization of CHemicals) in Europe and *Toxicity*

Testing for the 21st Century (National Academy of Sciences & National Research Council 2007) in the US.

Second, toxicity databases now contain large amounts of chemical and biological information through data consolidation (e.g. ACToR, TOXNET, DSSTox, (Judson et al. 2009)) and large-scale testing. Programs such ToxCast (Judson et al. 2010), Tox21 (Collins et al. 2008), Molecular Libraries Initiatives (Austin et al. 2004) perform high throughput screening (HTS) on thousands of chemicals over thousands of biological endpoints. The enlarged data scale in terms of broader chemical scope (chemical breadth) and deeper biological assay characterization (biological depth) has opened up new opportunities for cheminformatics and bioinformatics. Where previously only a few chemicals were tested, the broader chemical scope of the data has reinvigorated interest in cheminformatics to transform latent chemical patterns into useful chemical insight. On the other hand, the deeper biological assay characterization allows one to learn more about each chemical in terms of its biological responses. Yet, sticking to the approach of only chemical or biological modeling is unlikely to take full advantage of the richness of the data that may be unlocked by integrating the two.

Third, the many parallels between bioinformatics and cheminformatics provide points of commonality to facilitate integration. Underpinning both fields are statistical functions relating molecular features of a chemical to its behavior. These statistical relationships rely on the similarity principle which expects chemicals similar in their molecular feature profiles to exhibit similar behavior. The key difference between cheminformatics and bioinformatics here lies in the choice of appropriate molecular features, whether as ‘omics profiles assayed by HTS or molecular structural information represented by chemical descriptors. The statistical techniques, whether as simple as read-across or as complex as machine learning, are equally applicable to both fields.

As such, one possible means of integration is to apply existing statistical methods to both types of molecular features, chemical and biological (data pooling/integration, Figure 1.1). Another way is to merge chemical models with biological models (model pooling or ensemble modeling). Other approaches may be less straightforward, strategically combining chemical structures and biological assays such that the two data sources compensate for each other's shortcomings and the complementary information between them is maximally used.

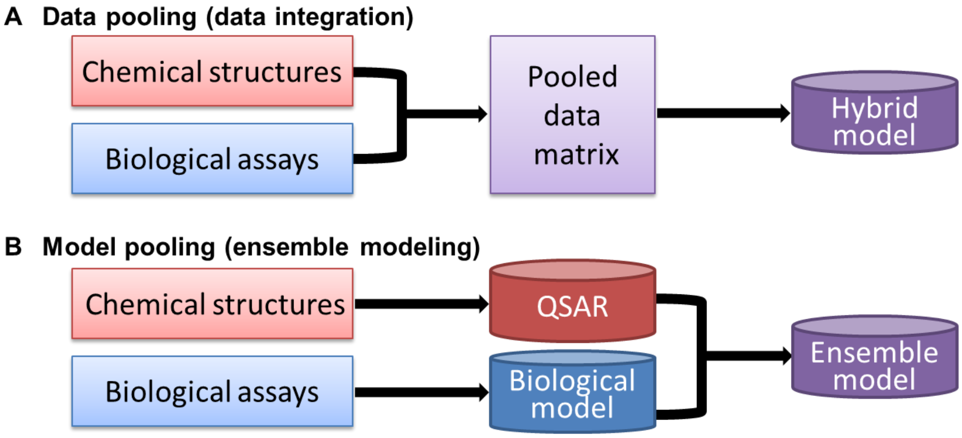


Figure 1.1: Integrative chemical-biological approaches for toxicity prediction

1.4. Review of integrative chemical-biological modeling for predictive toxicology

Using the classification scheme described above (Figure 1.1), recent integrative efforts merging chemical and biological data are reviewed below (Table 1.1). Because many of the integrative methods are not specific to a particular type of data, they may also incorporate other data types such as text annotations mined from biomedical literature or drug labels and clinical data from health databases.

Table 1.1: Integrative approaches used for toxicity prediction

Prediction target	Data sources	Integrative approach	Publication
Rat LD_{50}	Chemical structures, Cytotoxicity	Data pooling	(Zhu et al. 2008)
Rat LD_{50}	Chemical structures, Cytotoxicity	Other integrative method	(Zhu et al. 2009)
Rat LD_{50}	Chemical structures, Dose-cytotoxicity profiles	Data pooling	(Sedykh et al. 2011)
Rat reproductive toxicity	Chemical structures, *In vitro* assays	Other integrative method	(Zhang 2011)
In vivo toxicities	Chemical structures, *In vitro* assays	Data pooling	(Thomas et al. 2012)
Drug hepatotoxicity	Chemical structures, Transcriptomics	Data pooling, Model pooling, Other integrative method	Chapter 2 (Low et al. 2011) Chapter 3
Drug hepatotoxicity	Chemical structures, Hepatocyte imaging assays	Data pooling	(Zhu et al. 2013)
Adverse drug reactions	Chemical structures, Drug properties, Adverse drug reactions	Data pooling, Other integrative method	(Cami et al. 2011)
Adverse drug reactions	Chemical structures, Electronic health records	Model pooling	(Vilar et al. 2011, 2012)
Adverse drug reactions	Chemical structures, Bioactivities, Adverse drug reactions, Therapeutic indications	Data pooling	(Liu et al. 2012)
Adverse drug reactions	Chemical structures, Drug targets, Adverse drug reactions, Clinical outcomes	Data pooling, Other integrative method	(Cheng et al. 2012, 2013)
Adverse drug reactions	Chemical structures, Bioactivities	Other integrative method	(Yamanishi et al. 2012)
Drug properties	Chemical structures, Bioactivities	Other integrative method	(Lounkine et al. 2011)
Drug targets	Chemical structures, Adverse drug reactions	Data pooling, Other integrative method	(Campillos et al. 2008)
Drug targets	Chemical structures, Protein sequence	Data pooling, Other integrative method	(Yamanishi et al. 2008)
Drug targets	Chemical structures, Adverse drug reactions	Data pooling, Model pooling, Other integrative method	(Oprea et al. 2011)
Drug targets	Chemical structures, Adverse drug reactions	Data pooling, Other integrative method	(Lounkine et al. 2012)
Drug targets associated with agranulocytosis	Protein docking profiles Transcriptomics	Other integrative method	(Yang et al. 2011)

1.4.1. Chemical-biological data pooling (data integration)

One way of chemical-biological integration is data pooling or data integration in which disparate data sources are pooled to create a larger data matrix for modeling by existing statistical methods. This has been aided by the growing availability of public repositories such as PubChem, ChEMBL and ACToR/DSSTox. Besides high throughput experimentation, automated data generation has expanded non-traditional sources of data such as text annotations mined from biomedical literature (InSTEM, ChemoText), product labels (SIDER) and clinical notes (Bai & Abernethy 2013, Chiang & Butte 2009, Iskar et al. 2012, Oprea et al. 2007).

Table 1.1 includes several studies predicting toxicity from pooling various combinations of data. Generally, prediction performance improved as data were pooled. However, several exceptions exist. A comprehensive evaluation of models predicting 60 *in vivo* toxicities from chemical structures and/or *in vitro* assays in ToxCast phase I described mixed success with data pooling (Thomas et al. 2012). Zhu et al. reported lower predictivity of hepatotoxicity from data pooling of chemical structures and hepatocyte imaging profiles (Zhu et al. 2013).

To overcome the limited prediction performance from data pooling, additional data treatment may be required, especially when biological assay data include considerable experimental noise. For example, Sedykh et. al. introduced a noise filter to transform cytotoxicity profiles into dose-response curve parameters that, when pooled with chemical structures, provided more accurate models of rat acute toxicity than the original cytotoxicity assay values (Sedykh et al. 2011). Chapter 2 explores the pooling of chemical structures and toxicogenomics profiles for hepatotoxicity prediction and compares the resultant "hybrid" models against the chemical-only or biological-only models.

1.4.2. Chemical-biological model pooling (ensemble modeling)

Another way of integrating chemical and biological data is by ensemble modeling which pools individual predictions from several models into a final predicted value. The main benefit of ensemble modeling, increased predictivity, arises when the constituent models compensate for the errors of one another (Dietterich 2000). The notion that many models are better than one is best exemplified by the random forest algorithm which seeks the consensus vote of numerous constituent decision tree models within its "forest" (Breiman 2001). In the case of toxicity modeling, chemical-based models and biological-based models may be pooled such that their consensus vote provides the final prediction outcome.

Such model pooling is already widely practiced in regulatory assessment and drug discovery during which the user weighs all the prediction outcomes from various toxicity models before arriving at a consensus decision (Kruhlak et al. 2012, Wang et al. 2012). For example, drugs must not contain structural alerts of mutagenicity and their bioassay profiles must not inhibit the major cytochrome P450 enzymes required for drug metabolism.

Ensemble modeling can be used in one of two ways. One can require that all the constituent models for a compound point to the same prediction outcome such that their intersection represents an enriched space in which toxicity can be estimated with higher confidence. Alternatively, one can argue that ensemble modeling enlarges the modelable space of molecules such that compounds that cannot be predicted with confidence by one model be supplanted by another model that can. In the first case, Vilar et. al. showed increased precision when a chemical similarity model was pooled with a model based on clinical notes (Vilar et al. 2011, 2012). In the second case, as Chapter 3 will illustrate, the ensemble chemical-biological model compensates for the invalid predictions by the QSAR model outside the chemical coverage area. While conceptually simple, ensemble models may not always outperform their constituent chemical and biological models, as Chapter 3 will demonstrate on four data sets containing chemical structures and bioassays.

1.4.3. Other integrative chemical-biological modeling

The shortcomings of merely pooling data or models have led to more innovative integrative approaches that rely on the rational use of data and modeling methods. Zhu et. al. described a two-step hierarchical approach which first stratified compounds by their *in vitro-in vivo* correlation and then built stratum-specific models. Poorly correlated compounds with *in vitro* surrogates were assumed to be strongly influenced by biological factors and would benefit from the inclusion of biological data. Such compounds were shown to benefit from models pooling chemical structures and *in vitro* assay data (Zhang 2011). Such strategic use of biological data to stratify data sets into clusters for localized modeling was also attempted by Lounkine et al. who clustered compounds by chemical similarity and their bioactivity (Lounkine et al. 2011).

Another class of integrative approaches employ network modeling which allows the simultaneous study of disparate entities (chemicals, targets and phenotypes) (Berger & Iyengar 2009). In a network, entities (nodes) are connected (edges) if they are associated. Association may be defined in terms of physical interactions (*e.g.*, drug binds to target) or statistical associations. In such modeling, the goal is to discover new associations among the entities through indirect associations. This is best illustrated by Swanson’s ABC paradigm (Swanson 1986) in which entities A and C are indirectly associated if there exist direct associations between pairs A-B and B-C (Figure 1.2A). Networks may be further enriched by chemical similarity (Lounkine et al. 2012, Oprea et al. 2011), protein sequence similarity

(Yamanishi et al. 2008), side effect similarity (Campillos et al. 2008) such that novel inferences can be drawn (Figure 1.2B (Tatonetti et al. 2009))

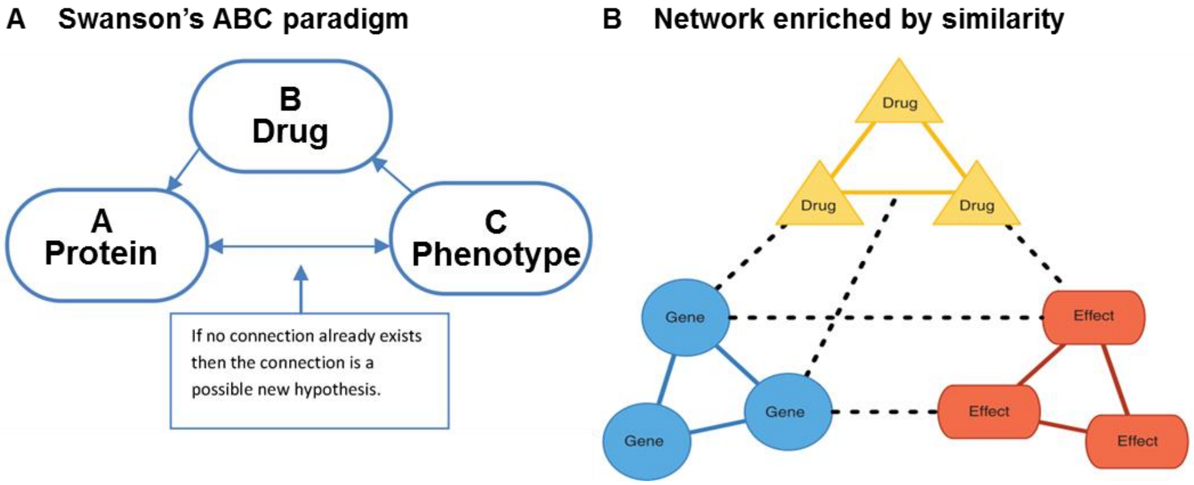

A Swanson's ABC paradigm
B
Drug
A
Protein
C
Phenotype
If no connection already exists then the connection is a possible new hypothesis.
B Network enriched by similarity
Drug
Drug
Drug
Gene
Gene
Gene
Effect
Effect
Effect

Figure 1.2: (A) Swanson ABC paradigm, adapted from (Baker & Hemminger 2010) (B) Network enriched by similarity (solid edges) enable novel inferences (dotted edges) to be drawn, adapted from (Tatonetti et al. 2009)

Associations successfully predicted in recent studies include those of target-phenotype (Lounkine et al. 2012), chemical-phenotype (Cami et al. 2011, Cheng et al. 2013), chemical-target (Campillos et al. 2008, Yang et al. 2011). For examples of broader efforts to infer more than a single type of associations, readers are best referred to (Berger & Iyengar 2009, Cheng et al. 2012, Oprea et al. 2011, Tatonetti et al. 2012)

Another integrative method, quantitative chemical-biological read-across (CBRA) based on the principles of *k* nearest neighbors, is presented in Chapter 3. Unlike an ensemble model that pools chemical-based predictions and biological-based predictions, enhanced pooling utilizing similarity weights maximizes the complementarity between chemical and biological data and resolve their conflicting predictions. Chapter 3 also compares the three types of integrative approaches (data pooling, model pooling and CBRA) on four data sets.

1.5. Human health data and epidemiology in predictive toxicology

Besides chemical and bioassay databases, increasing digitization of health databases (*e.g.* health insurance claims, national health records) offers new ways of studying toxic health effects in human populations (Adami et al. 2011, Hall et al. 2012). Health data may be more informative than toxicology studies performed in non-human model organisms which sometimes extrapolate poorly to humans due to inherent interspecies differences. To bridge the disconnect, one can draw upon human health data and apply epidemiological methods to systematically study health effects in human populations.

Epidemiology and toxicology have always complemented each other: epidemiology provides the tools for discovery, to reliably identify the risk factors of a certain health outcome while toxicology provides the tools for corroboration, to verify that the risk factors are indeed causative through experiments and to suggest a plausible mode of action. Sometimes, their roles reverse. Toxicology, taking on a predictive role, may accumulate reasonable experimental evidence from non-human studies to suspect a chemical of human toxicity. Epidemiology, now taking on a confirmatory role, attempts to verify the toxicological findings in humans (Adami et al. 2011).

Drug safety, in particular, could benefit from the integration of toxicology and epidemiology, specifically pharmacoepidemiology, the specific branch of epidemiology concerned with the study of drug effects in human populations. Although human drug toxicity is extensively investigated in clinical trials prior to market approval, many adverse drug reactions (ADR) may have been missed in clinical trials which do not reflect the "real-world" setting with pediatric or geriatric patients or patients with co-morbidities (Arellano 2005, Strom et al. 2012). ADR are a major cause of medical errors and account for $75 billion of unnecessarily healthcare expenditure (Ahmad 2003, National Research Council 2007).

Current ADR detection by most drug authorities (e.g. US Adverse Events Reporting System, World Health Organization VigiBase) relies on the passive surveillance of spontaneous ADR reports. When an unusually high number of ADR reports are linked to a drug, determined by statistical tests of association, a warning signal is generated (Bate & Evans 2009, Harpaz et al. 2012). However, passive surveillance systems are prone to underreporting and reporting bias as the spontaneous reports are only voluntary for healthcare professionals and patients.

In response, recent initiatives such as the Food and Drug Administration (FDA) Sentinel Initiative (Platt et al. 2012) and EU-ADR (Oliveira et al. 2012) have called for active surveillance of ADR. Such a system, instead of passively relying on spontaneous reports, actively monitors clinical data for ADR signals. Increasingly, ADR prediction is performed with alternative data sources such as patient health records (LePendu et al. 2013), health administrative databases, patient web forums (White et al. 2013), biomedical literature (Bisgin et al. 2011, Shetty & Dalal 2011), chemical structures (Bender et al. 2007, Matthews et al. 2009a,b; Scheiber et al. 2009) and bioassays (Chiang & Butte 2009, Pouliot et al. 2011).

Because of the large amount of data involved, effective surveillance aims for a tiered approach in which signal detection composes of three stages: signal generation, refinement and evaluation (Platt et al. 2012). In the first stage (signal generation), high throughput data mining methods generate suspect drug-ADR pairs. The second stage (signal refinement) typically employs pharmacoepidemiology to check if the initial signal persists after statistical adjustment for confounders such as demography, co-medications and co-morbidities (McClure et al. 2012). Signals progressing to the third stage (signal evaluation) will then undergo a careful clinical expert evaluation.

Therefore, one practical way for implementing an effective ADR detection system is to first apply high throughput cheminformatics techniques which utilize readily available drug chemical structures for the prediction of potential ADR (signal generation). Then, the potential ADR are assessed by pharmacoepidemiology using patient health data (signal refinement). As pharmacoepidemiology calls for thoughtful study design and rigorous statistical analysis, it, unlike cheminformatics, is less amenable to large-scale automated analysis. Coupling cheminformatics to pharmacoepidemiology will combine the best of both methods: the high throughput advantage of the former and the statistical rigor of the latter. Such an approach is exemplified in Chapter 4 for the prediction and validation of drugs inducing Stevens Johnson Syndrome (SJS), an ADR of major concern.

1.6. Dissertation outline

This dissertation presents integrative approaches addressing some of the above problems facing predictive toxicology. Poor model performance due to the lack of biological factors in cheminformatics and chemical structures in bioinformatics may be overcome by integrative modeling of both chemical and biological factors. Chapter 2 illustrates hepatotoxicity prediction from the combined use of chemical structures and toxicogenomics assays with existing machine learning methods [*k* nearest neighbors (*k*NN), support vector machines (SVM), random forests (RF) and distance-weighted discrimination (DWD)].

Chapter 3 extends the work in Chapter 2 by developing a new integrative method, chemical-biological read across (CBRA), that exploits the complementary information between chemical structure and bioassays for more accurate prediction.

Chapter 4 attempts to account for human health effects in the study of ADR through the use of pharmacoepidemiology to evaluate human health data. Compared to current ADR detection methods which rely only on spontaneous ADR reports, Chapter 4 draws from various data sources (chemical structures, spontaneous ADR reports, health insurance claims) and various methods (cheminformations, pharmacoepidemiology) for the study of Stevens Johnson Syndrome ADR. It provides a feasible workflow coupling high throughput cheminformatics with in-depth pharmacoepidemiology analysis, a process in which cheminformatics predicts high risk drugs for pharmacoepidemiology validation.

Chapter 4 also addresses the lack of interpretability of QSAR “black box” modeling by proposing an interpretation framework for identifying important chemical substructures associated with SJS.

The concluding chapter interweaves the findings from chapters 2 to 4 and discusses their contributions towards predictive toxicology. Chapter 5 also examines their study limitations, some of which may provide the motivation for future research.

CHAPTER 2. INTEGRATIVE CHEMICAL-BIOLOGICAL MODELING WITH EXISTING METHODS: PREDICTING DRUG-INDUCED HEPATOTOXICITY USING QSAR AND TOXICOGENOMICS APPROACHES[1]

2.1. Overview

Quantitative structure-activity relationship (QSAR) modeling and toxicogenomics are typically used independently as predictive tools in toxicology. In this study, we evaluated the power of several statistical models for predicting drug hepatotoxicity in rats using different descriptors of drug molecules, namely, their chemical descriptors and toxicogenomics profiles. The records were taken from the Toxicogenomics Project rat liver microarray database containing information on 127 drugs (http://toxico.nibio.go.jp/datalist.html). The model end point was hepatotoxicity in the rat following 28 days of continuous exposure, established by liver histopathology and serum chemistry. First, we developed multiple conventional QSAR classification models using a comprehensive set of chemical descriptors and several classification methods (*k* nearest neighbor, support vector machines, random forests, and distance weighted discrimination). With chemical descriptors alone, external predictivity (correct classification rate, CCR) from 5-fold external cross-validation was 61%. Next, the same classification methods were employed to build models using only toxicogenomics data (24 h after a single exposure) treated as biological descriptors. The optimized models used only 85 selected toxicogenomics descriptors and had CCR as high as 76%. Finally, hybrid models combining both chemical descriptors and transcripts were

[1] This chapter is a reproduction, in whole, with permission of a publication in Chemical Research in Toxicology by (Low et al. 2011) at doi:10.1021/tx200148a.

developed; their CCRs were between 68 and 77%. Although the accuracy of hybrid models did not exceed that of the models based on toxicogenomics data alone, the use of both chemical and biological descriptors enriched the interpretation of the models. In addition to finding 85 transcripts that were predictive and highly relevant to the mechanisms of drug-induced liver injury, chemical structural alerts for hepatotoxicity were identified. These results suggest that concurrent exploration of the chemical features and acute treatment-induced changes in transcript levels will both enrich the mechanistic understanding of subchronic liver injury and afford models capable of accurate prediction of hepatotoxicity from chemical structure and short-term assay results.

2.2. Introduction

Hepatotoxicity is a major factor contributing to the high attrition rate of drugs. At least a quarter of the drugs are prematurely terminated or withdrawn from the market due to liver-related liabilities (Schuster et al. 2005). As a result, modern drug development has evolved into a complex process relying on the iterative evaluation of multiple data sources to eliminate potentially harmful candidates as cheaply and as early as possible. In addition, high throughput, high content, and other data-rich experimental techniques, accompanied by the appropriate informatics tools, are rapidly incorporated into toxicity testing.

Quantitative structure–activity relationship (QSAR) modeling is widely used as a computational tool that allows one to relate the potential activity (*e.g.*, toxicity) of an agent to its structural features represented by multiple chemical descriptors. As with any multivariate statistical modeling, rigorous validation procedures are necessary to guard against overfitting and overestimating model predictivity (Tropsha 2010). QSAR models have demonstrated good predictivity especially for specific end points such as solubility or binding affinity to a certain target. However, QSAR predictivity is generally poor in the case of a complex end point such as hepatotoxicity where the structure–activity relationship is less straightforward due to multiple mechanisms of action (Hou & Wang 2008).

Toxicogenomics is now routinely used in drug and chemical safety evaluation, providing valuable mechanistic understanding of the molecular changes associated with the disease or treatment (Cui & Paules 2010). In addition, its utility for predicting toxicity has been explored. Blomme et al. developed models using transcriptional changes after short-term (5 days) exposure to predict bile duct hyperplasia that otherwise required long-term *in vivo* experiments (Blomme et al. 2009). Fielden et al. developed a 37-gene classification model using microarray data following short-term (1–5 days) exposure to predict nongenotoxic hepatocarcinogenicity with over 80% accuracy (Fielden et al. 2007). Zidek et al. reported high accuracy with a 64-gene classifier for the prediction of acute hepatotoxicity (Zidek et al. 2007). The Toxicogenomics Project in Japan, set up by the Ministry of Health, Labour and Welfare, National Institute of Health Sciences, and 15 pharmaceutical companies, has also identified several toxicogenomics signatures indicative of the various toxic modes of action such as phospholipidosis (Hirode et al. 2008), glutathione depletion (Kiyosawa et al. 2007), bilirubin elevation (Hirode et al. 2009a), nongenotoxic hepatocarcinogenesis (Uehara et al. 2008), and peroxisome proliferation (Tamura et al. 2006).

Most previous studies on statistical modeling of toxicity used either chemical descriptors (conventional QSAR) or toxicogenomics profiles independently for model development. However, in our recent studies, we have demonstrated the benefits of hybrid classification models of *in vivo* carcinogenicity (Zhu et al. 2008) and toxicity (Sedykh et al.

2011), and employing both chemical descriptors and biological assay data (treated as biological descriptors). In the first study of this type (Zhu et al. 2008), we used the results of high-throughput screening assays of environmental chemicals along with their chemical descriptors to arrive at improved models of rat carcinogenicity. This approach was extended to predicting acute toxicity half-maximal lethal dose in rats using dose–response *in vitro* data as quantitative biological descriptors (Sedykh et al. 2011).

Following our hybrid (chemical and biological descriptors) data modeling paradigm, we sought to integrate QSAR and toxicogenomics data to develop classification models of hepatotoxicity using a data set of 127 drugs studied in the Japanese Toxicogenomics Project (Uehara et al. 2010). We built classifiers combining chemical descriptors and toxicogenomics data alongside the conventional QSAR, as well as toxicogenomics models. Our objective was to investigate if chemical descriptors and biological descriptors, such as gene expression, could be complementary. In addition, we sought to enhance the interpretation of the models in terms of elucidating the chemical structural features and biological mechanisms associated with hepatotoxicity. We show that statistically significant and externally predictive models can be developed by combining chemical and biological descriptors and can be used to predict hepatotoxicity and prioritize chemicals for toxicogenomics and other *in vivo* studies.

2.3. Materials and Methods

2.3.1 Data

The chemical name, dosage, administration route, and vehicle for the 127 compounds used in this study are summarized in Appendix 1 Table A1.2.1. The detailed protocol for the animal study was described previously (Uehara et al. 2010). Briefly, 6-week old male Sprague–Dawley rats (Charles River Japan, Inc., Kanagawa, Japan) with five animals per group were used in the study. Animals were sacrificed 24 h after a single dose or 24 h after repeat daily treatment for 28 days. Blood samples were collected from the abdominal aorta under ether anesthesia. Serum chemical indicators included alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, total bilirubin, direct bilirubin, and gamma-glutamyl transpeptidase. The livers were quickly removed following exsanguination and sections of the livers were placed in 10% phosphate-buffered formalin for histopathology. Formalin-fixed liver tissue was embedded in paraffin, and sections were stained with hematoxylin and eosin and examined histopathologically under light microscopy. Remaining liver tissues from left lateral lobes were soaked in RNALater (Ambion Inc., Austin, TX) and stored at −80 °C until used for microarray analysis. Detailed methods for microarray analysis were previously reported (Uehara et al. 2010). Raw microarray data files with individual animal histopathological data are available (http://toxico.nibio.go.jp/datalist.html). In this study, toxicogenomics data obtained from rats treated with a single dose of a drug or vehicle for 24 h was used. The experimental protocols were reviewed and approved by the Ethics Review Committee for Animal Experimentation of the National Institute of Health Sciences (Tokyo, Japan).

Liver histopathology and serum chemistry in animals treated for 28 days were assessed for the determination of the hepatotoxicity end point for prediction. Histopathology was graded by two trained pathologists in a blinded manner as follows: no change, very slight (minimal), slight, moderate, and severe. Spontaneously observed lesions (*e.g.*, minimal focal necrosis and microgranuloma) were not used for grading. The results of a histopathology analysis were considered positive if the grade recorded was other than "no change." Appendix 1 Table A1.2.1 lists serum chemistry and histopathology classification for each compound. A compound was denoted hepatotoxic if it exhibited histopathology characteristics of hepatotoxicity (*e.g.*, hepatocellular necrosis/degeneration, inflammatory cell infiltration, bile duct proliferation, etc.) regardless of the findings from serum chemistry. Conversely, a compound was deemed nonhepatotoxic if it did not result in adverse histopathological features. When the histopathological observations were inconclusive (*e.g.*, hepatocellular hypertrophy, vacuolization, etc.), serum chemistry data was considered. Under these circumstances, significant changes (Dunnett's test) in at least one enzyme marker would render the compound hepatotoxic. Otherwise, the compounds with inconclusive histopathology and normal serum chemistry were denoted nonhepatotoxic. In total, there were 53 (42%) hepatotoxic and 74 (58%) nonhepatotoxic compounds.

2.3.2. Curation of chemical data

The data set was curated according to the procedures described by (Fourches et al. 2010). Briefly, counterions and duplicates were removed, and specific chemotypes such as aromatic and nitro groups were normalized using several cheminformatics software such as ChemAxon Standardizer (v.5.3, ChemAxon, Budapest, Hungary), HiT QSAR (Kuz'min et al. 2008) and ISIDA (Varnek et al. 2008). Following the automated curation, the data set was inspected manually, and two metal-containing compounds for which most chemical descriptors cannot be calculated, cisplatin and carboplatin, were removed. Chemical descriptors were calculated with Dragon (v.5.5, Talete SRL, Milan, Italy) and Molecular Operating Environment (MOE, v.2009.10, Chemical Computing Group, Montreal, Canada) software. Simplex representation of molecular structure (SiRMS) descriptors were derived as detailed elsewhere (Muratov et al. 2010). After range scaling (from 0 to 1), low variance (SD $< 10^{-6}$) and highly correlated descriptors (if pairwise $r^2 > 0.9$, one of the pair was randomly removed) were removed. QSAR models were built separately using 304 Dragon, or 116 MOE, or 271 SiRMS descriptors (Figure 2.1).

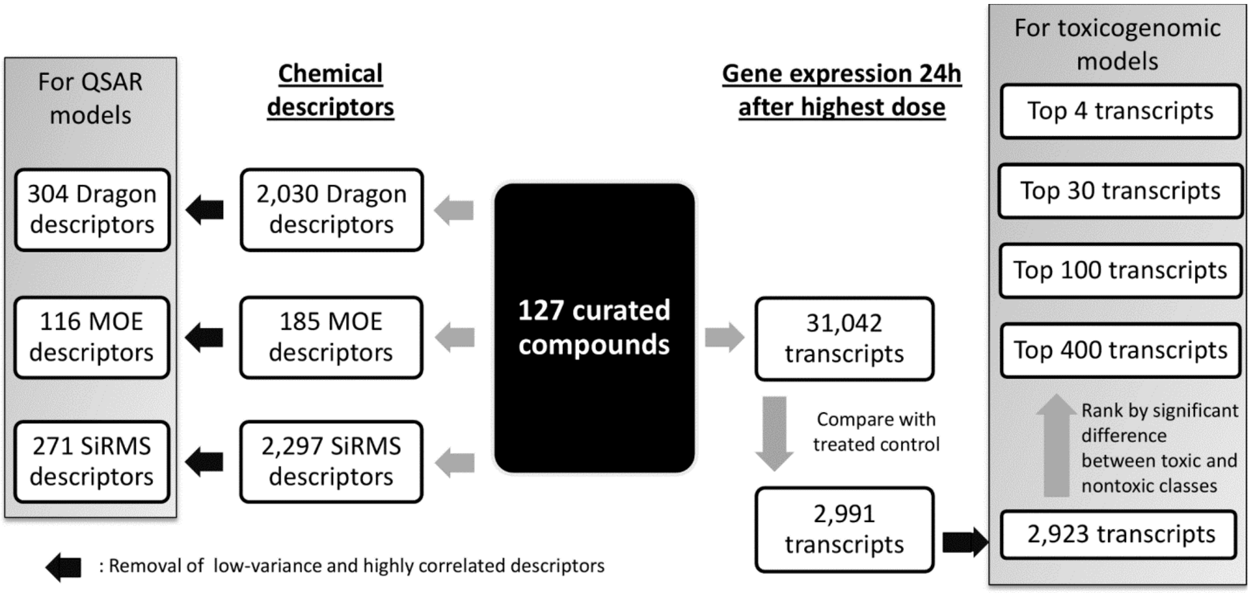


Figure 2.1. Workflow illustrating data curation and feature selection for modeling.

2.3.3. Selection of transcripts

Transcripts were selected for modeling using various feature selection methods. Of the 31,042 transcripts measured, we removed those consistently absent across all compounds. Then we extracted 2,991 transcripts with sufficient variation across all the compounds on the basis of the following criteria: the largest change of any transcript over its untreated equivalent must exceed 1.5-fold, and the smallest false discovery rate (Welch t-test) must be less than 0.05. Next, transcripts with low variance (all, or all but one value is constant) and high correlation (if pairwise $r^2 > 0.9$, one of the pair, chosen randomly, was removed) were excluded leaving 2,923 transcript variables (Figure 2.1) which were range scaled.

Then, supervised selection methods were used to filter genes differentially expressed between hepatotoxic and nonhepatotoxic compounds. Significance analysis of microarrays (SAM) (Tusher et al. 2001), a permutation variant of the t-test commonly used for transcript selection, was used. Top ranked transcripts were retained for modeling. Different sets of transcripts were selected for each modeling set used in 5-fold external cross-validation to avoid selection bias introduced by a supervised selection process.

2.3.4. Modeling and Validation

*k*NN (Zheng & Tropsha 2000), SVM (Vapnik 2000), random forest (RF) (Polishchuk et al. 2009), and distance weighted discrimination (DWD) (Marron et al. 2007) machine learning techniques, designed to effectively handle high dimension-low sample size data, were used for modeling. The modeling workflow (Tropsha 2010, Tropsha & Golbraikh 2007) used both internal and external validation (Appendix 2 Figure A2.2.1). In a 5-fold external cross-validation, 127 compounds were randomly partitioned into 5 subsets of nearly equal size. Each subset was paired with the remaining 80% of the compounds to form a pair of external and modeling sets. The data within each modeling set were further divided into multiple pairs of training and test sets for internal validation.

Although models were built using the training set, model selection depended on their performance on both the training and test sets (*i.e.*, internal validation) since training set accuracy alone is insufficient to establish robust and externally predictive models (Golbraikh & Tropsha 2002b). The prediction outcome for each model was categorized as “0” for nontoxic compounds or “1” for toxic ones. Selected models were then pooled into a consensus model by simple averaging and used to predict the hepatotoxicity of compounds in the external sets (*i.e.,* external validation). The toxicity threshold was set at 0.5 unless otherwise mentioned, *i.e.,* a compound is predicted to be nontoxic if a consensus mean is less than 0.5 and toxic otherwise.

The *y*-randomization test was employed to ensure that there was no chance correlation between selected descriptors and hepatotoxicity. After random permutation of the hepatotoxicity labels in the modeling sets, models were rebuilt following the same workflow, and their CCR values for both training and test sets were collected and compared. This test was repeated at least three times. Models generated from the randomized labels were expected to perform significantly worse than those derived from the original data set.

All reported model predictivity measures, specificity, sensitivity, and correct classification rate, were obtained from 5-fold external cross-validation. Specificity denotes the true negative rate, or the rate correctly predicted within the nonhepatotoxic class. Similarly, sensitivity, the true positive rate, measures the rate correctly predicted within the hepatotoxic class. CCR is the average of the rates correctly predicted within each class (CCR = [specificity + sensitivity]/2). Coverage is the percentage of compounds in the external set within the applicability domain (AD) of the model. The AD is a similarity threshold within which compounds can be reliably predicted (Tropsha et al. 2003).

Chemical and toxicogenomics descriptors found to be predictive were subsequently analyzed. Ingenuity Pathway Analysis (Ingenuity Systems, Redwood City, CA) software was used for the functional analysis of the significant transcripts. The networks were constructed on the basis of predefined molecular interactions in the Ingenuity database, and the Ingenuity score was used to rank pathways for analysis. Chemicals were clustered by the selected toxicogenomics descriptors using an unsupervised self-organizing map (SOM) in R (Kohonen package). Chemical structural alerts for hepatotoxicity were identified using HiT QSAR (Kuz'min et al. 2008) and verified with XCHEM (Sedykh & Klopman 2006). Briefly, XCHEM searches for common structural motifs within each class and ranks them by their relative frequencies.

2.4. Results

2.4.1. Model development

First, we developed QSAR models of subchronic (28 days of treatment) hepatotoxicity using various types of chemical descriptors (Table 2.1). Prediction performance was generally poor (55–61% CCR) across all descriptor types and classification methods. Three compounds (tannic acid, vancomycin, and cyclosporine) with molecular weights exceeding 1,200 (median molecular weight of the data set was 285) were excluded from the data set, corresponding to a coverage of 98% for some of the models. Given the generally unpromising results of the QSAR models described in Table 2.1, further combinatorial-QSAR (Kovatcheva et al. 2004) efforts to systematically combine each descriptor type with each classification method were not attempted.

Table 2.1. 5-Fold External Cross-Validation Prediction Performance of QSAR Models

Descriptors	Dragon	Dragon	MOE	SiRMS
Method	*k*NN	SVM	*k*NN	RF
Specificity ± SD[a]	0.62±0.17	0.62±0.16	0.60±0.18	0.77±0.08
Sensitivity ± SD	0.56±0.14	0.48±0.17	0.56±0.16	0.45±0.14
CCR ± SD	0.59±0.11	0.55±0.09	0.58±0.12	0.61±0.10
Coverage (%)	98	98	98	100

[a] SD refers to the standard deviation of the external predictivity measures (e.g. specificity) across the 5 folds.

Second, we developed classification models of subchronic (28 days of treatment) hepatotoxicity using liver toxicogenomics data obtained after a single dose treatment as a predictor of future toxicity. To find the optimal number of variables (transcripts), several sets of top ranking transcripts were selected (based on SAM analysis) for modeling by SVM, and the outcomes were compared (Figure 2.2). CCR ranged from 72% with top 4 significant transcripts per modeling fold to 78% with all 2,923 significant transcripts. An optimal model with a CCR of 76% was achieved when 30 transcripts per fold were used. These 5 sets of 30 transcripts per fold comprised of 85 unique transcripts across all folds, which may serve as predictive biomarkers (Appendix 1 Table A1.2.2). We used these 85 transcripts to develop additional models employing other classification methods (Table 2.2). The RF model had the highest performance with a CCR of 76%. DWD was also applied to the full set of 2,923 transcripts and had a CCR of 73%. The difference in performance between the QSAR and the toxicogenomic models was significant (p < 0.001).

Table 2.2. 5-Fold External Cross-Validation Prediction Performance of Toxicogenomics Models Based on the 85 Selected Transcripts

Method	*k*NN	SVM	DWD	RF
Specificity ± SD	0.82±0.08	0.84±0.10	0.77±0.11	0.84±0.05
Sensitivity ± SD	0.57±0.07	0.67±0.12	0.62±0.17	0.66±0.20
CCR ± SD	0.70±0.06	0.76±0.09	0.69±0.11	0.76±0.10
Coverage (%)	95	99	99	100

Third, we developed hybrid models of subchronic (28 days of treatment) hepatotoxicity using both chemical descriptors and single-dose treatment toxicogenomics data as biological descriptors. We studied how SVM model predictivity was affected when both the number of chemical descriptors and the number of transcripts were varied. To that effect, SAM was applied to independently rank chemical descriptors and transcripts, after which, different portions of top ranked variables were used for SVM modeling. Figure 2.2 shows that the CCR of the hybrid models did not exceed that of the models based on toxicogenomics data alone. However, hybrid models identified both important chemical descriptors and transcripts for the enhanced interpretation of the modeling outcomes. We could not have reliably detected the important chemical features from the relatively poorly fitted QSAR models. Adding transcripts boosted the predictivity of the hybrid models such that important chemical features were identified with greater confidence. Specifically, contributions of SiRMS descriptors used in RF hybrid models were interpreted using the approach of (Kuz'min et al. 2011) to uncover chemical substructures critical to hepatotoxicity. The substructures obtained through this analysis were compared to the alerts derived using XCHEM (Sedykh & Klopman 2006) and found to be concordant. The largest and most frequent substructures within each toxicity class are listed in Table 2.3 and provide evidence of the structure–activity relationship in the hybrid model. All QSAR, toxicogenomics, and hybrid models were significantly better than *y*-randomized models (p < 0.05 by Z-test), indicating that our models were not the result of chance correlations.

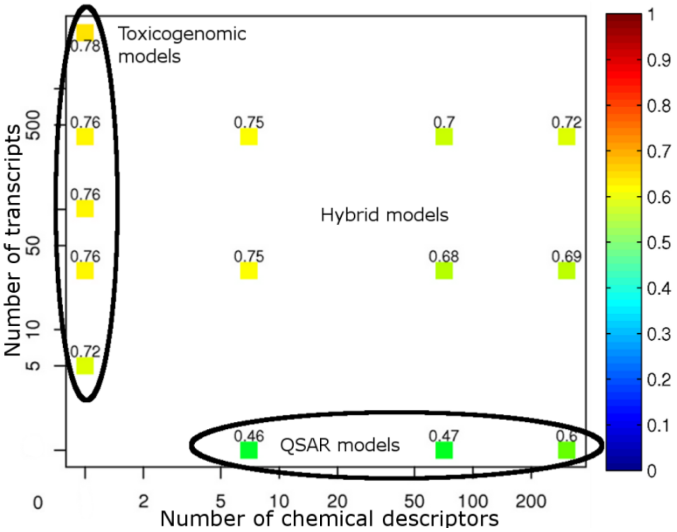
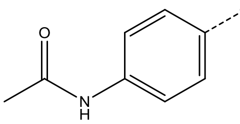
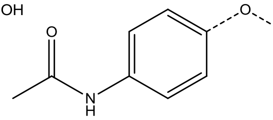
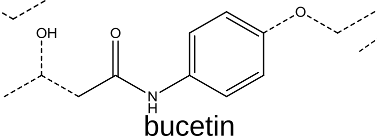
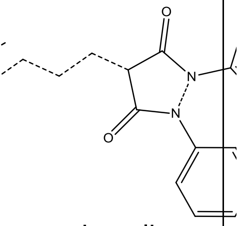
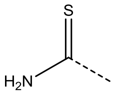
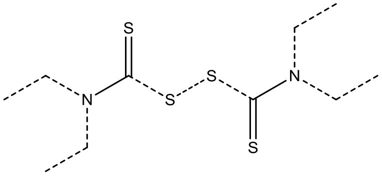
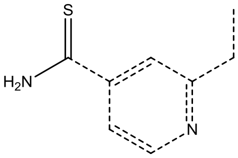
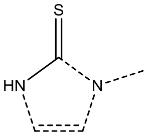
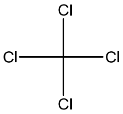
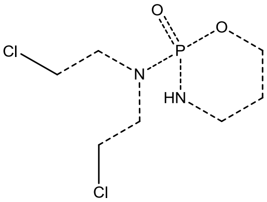
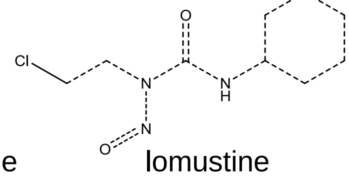
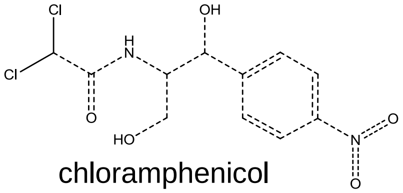
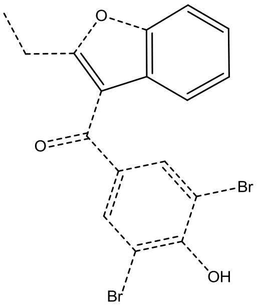
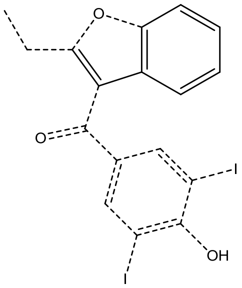
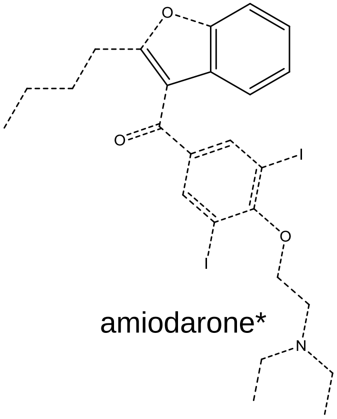
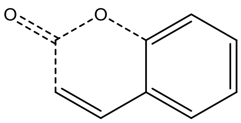


Figure 2.2. CCR accuracy of the models with respect to the number of chemical descriptors and transcripts used. All models were generated by SVM classification with 5-fold external cross-validation.

Table 2.3. Structural Alerts Mapped onto Example Compounds

Substructure A (Acetanilide)
 acetaminophen  phenacetin  bucetin  phenylbutazo
Substructure B (thioamide)
 thioacetamide  disulfiram  ethionamide  methimazole*
Substructure C (C-Cl)
 carbon tetrachloride  cyclophosphamide  lomustine  chloramphenicol
Substructure D (Styerene)
 benzbromarone  benziodarone  amiodarone*  coumarin

The toxicity threshold of the consensus models was set to 0.5, below which the compounds were classified as nontoxic and above which they were classified as toxic. Because the compounds on the margin are typically predicted with less confidence, we sought to determine the effect of adjusting the toxicity threshold on prediction performance. Figure 2.3A shows the distribution of QSAR-predicted values (using *k*NN method) for nontoxic and toxic compounds. Overall, the separation was poor due to a large proportion of nontoxic compounds that were predicted as toxic. While alternative thresholds yielding models with very high CCR may be selected (Figure 2.3C), severely reduced coverage of such models is a considerable drawback (Figure 2.3E). For example, setting two thresholds (dashed lines in Figure 2.3A), one at 0.36 (<0.36 are assigned nontoxic) and the second one at 0.56 (>0.56 are assigned as toxic) increased CCR to 68%, as compared to 59% with a single threshold of 0.5. However, the coverage of such a model was only 80% because the compounds whose predicted activities were between 0.36 and 0.56 could no longer be classified. Conversely, the toxicogenomics model developed with *k*NN showed good separation between toxic and nontoxic compounds (Figure 2.3B). A change in thresholds had a minor effect on the model's CCR and coverage (Figure 2.3D and 2.3F), showing that a single threshold was sufficient and that optimization of the activity thresholds would not be necessary. The optimal thresholds will be useful in the prediction of additional external compounds.

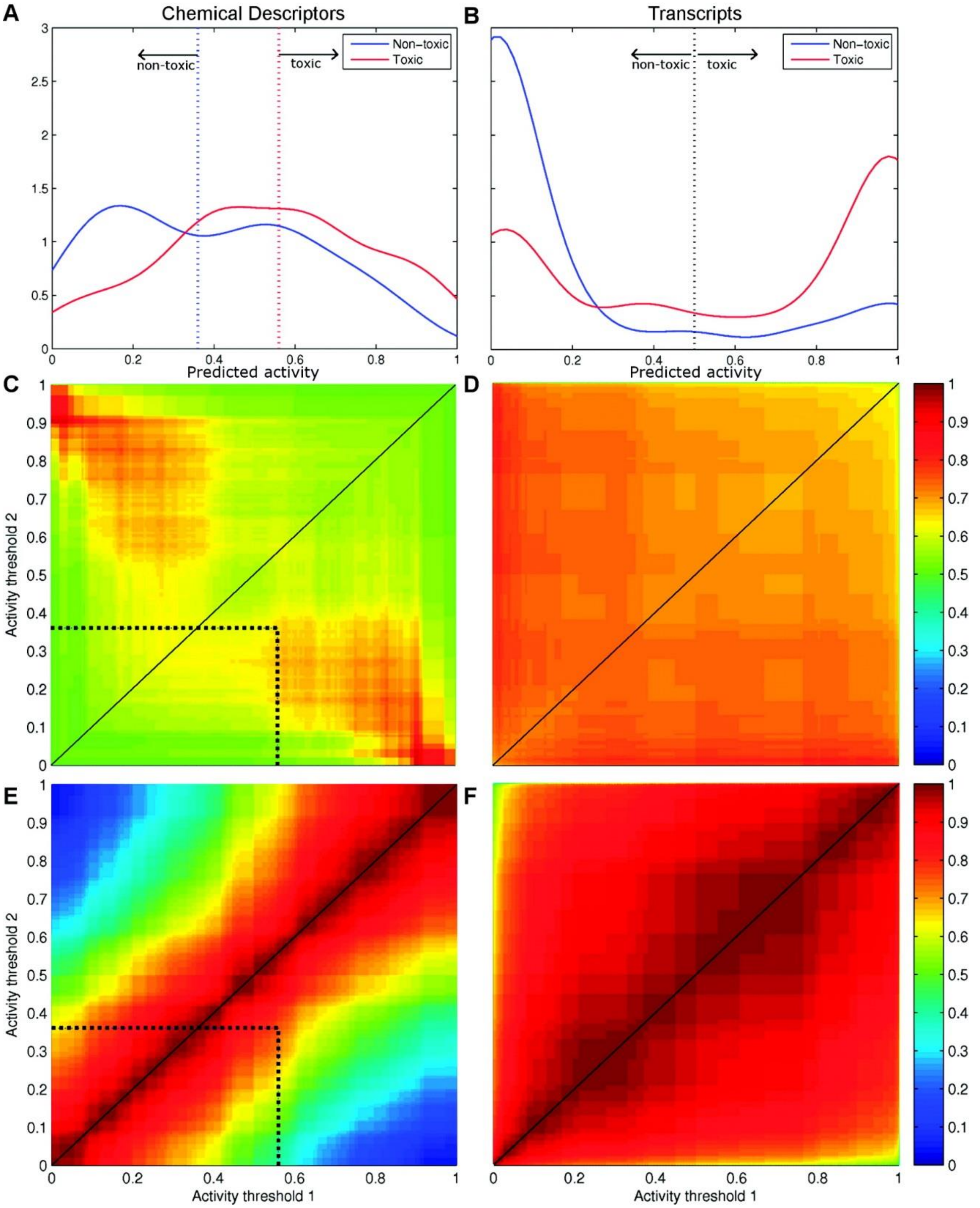


Figure 2.3. External prediction results of the QSAR (A, C, and E) and toxicogenomics (B, D, and F) models by *k*NN using different classification criteria. Classification accuracy (C and D, CCR) and coverage (E and F, percent chemicals within the applicability domain) results are shown. Dashed (A and B) and diagonal (B–F) lines denote a default single-threshold classification (threshold = 0.5). An example of a double-threshold classification (nontoxic if activity <0.36; toxic if activity >0.56) is shown by the dashed lines (C and E).

2.4.2. Model interpretation

Toxicogenomics data-based models were the most predictive of hepatotoxicity. To explore the biological significance and the mechanistic relevance of the selected 85 transcripts (64 up-regulated and 21 down-regulated), functional pathway analysis was performed. Hepatic nuclear factor 4α (Hnf4a)- and v-myc myelocytomatosis viral oncogene homologue (Myc)-centered interactomes were the two highest ranked networks involving large numbers of the 64 selected up-regulated genes (Appendix 2 Figure A2.2.4A–B and Table 2.IIIa). Canonical pathway analysis revealed that the eukaryotic initiation factor (*Eif*) 2 signaling pathway responsible for protein translation was up-regulated (Appendix 1 Table A1.2.IIIb). Among the down-regulated genes, the network involving cellular function and maintenance including transporters and inflammatory responses was the highest ranked network (Appendix 2 Figure A2.2.4C and Appendix 1 Table A1.2.IIIc). Canonical pathway analysis also revealed that many down-regulated genes were involved in the complement pathway (Appendix 1 Table A1.2.IIId).

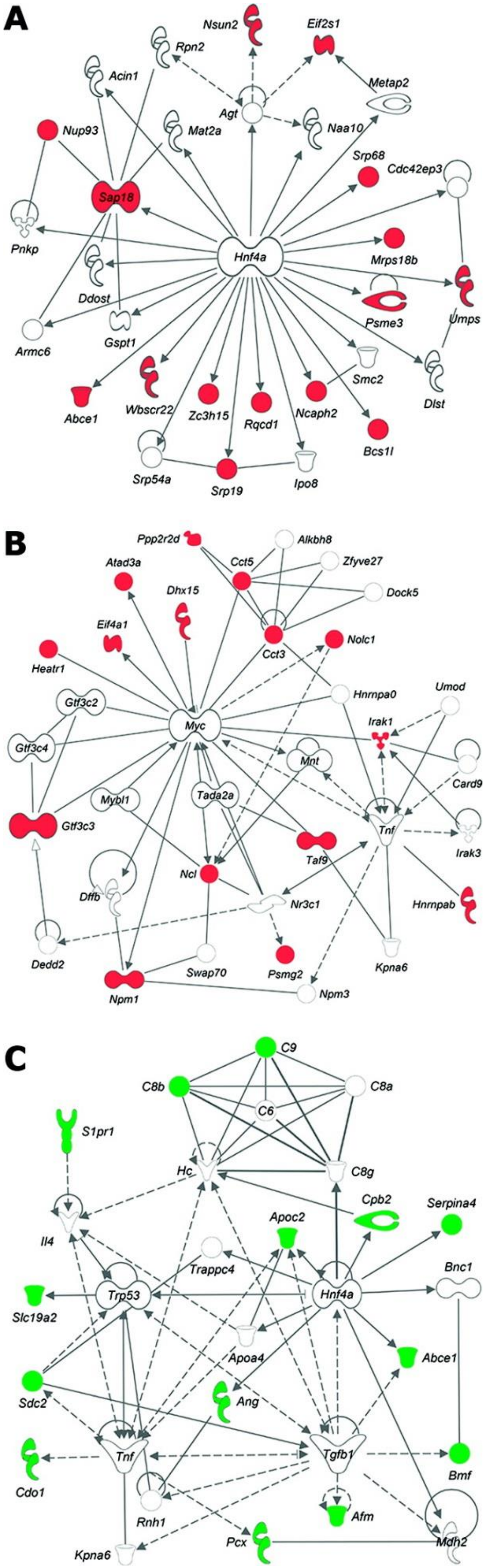


Figure 2.4. Molecular networks representing the toxicogenomics predictors of hepatotoxicity.

Hnf4a-centered (A), Myc-centered (B), and cellular function, and maintenance-related (C) interactomes were selected as the highest ranked networks among the 64 up- or 21 down-regulated genes used in modeling.

Red and green represent molecules up-regulated or down-regulated, respectively, by the hepatotoxic compounds. Ellipses, squares, triangles, trapezoids, lozenges, and circles represent transcription regulator, cytokine, kinase, transporter, enzyme, and other molecules, respectively. Arrows indicate molecular interactions, while lines indicate binding. Dashed arrows or lines indicate indirect interactions or binding.

See Tables 3.IIIa-d in the Appendix for a complete list of networks.

In addition, we used an unsupervised self-organizing map to cluster chemicals on the basis of their gene expression profiles (Figure 2.5 and Appendix 2 Figure A2.2.2). The objective was to uncover commonalities within clusters with similar gene expression profiles. As expected, the nonhepatotoxic agents were tightly clustered (green background). Among the hepatotoxic drugs (orange background), there were several clusters of compounds which may act through similar mechanisms of action. For example, oxidative stress-inducing agents (red text) such as acetaminophen, methapyriline, and nimesulide, and peroxisome proliferator-activated alpha (PPARα) agonists (blue text) such as fenofibrate, WY-14643, benzbromarone, clofibrate, and gemfibrozil formed two subclusters among the hepatotoxicants. The model-selected 85 transcripts were sufficient to cluster the drugs into toxicologically meaningful groups with similar modes of hepatotoxicity.

Understanding this difference in performance between the QSAR and the toxicogenomics models warrants an in-depth examination of the spatial distribution of compounds in their chemical and toxicogenomics descriptor space. Principal component analysis of the chemical features (Dragon descriptors, Figure 2.6A) and toxicogenomics data (85 selected transcripts, Figure 2.6B) demonstrated that the separation between nontoxic and toxic classes was poor in the chemical space. Appendix 1 Table A1.2.IVa lists 40 most chemically similar pairs of compounds. Half of them had opposite toxicities. Conversely, among pairs of compounds with the most similar gene expression profiles, only 23% exhibited opposite toxicities (Appendix 1 Table A1.2.IVb). In other words, pairs of compounds with similar gene expression profiles were more likely to have the same hepatotoxicity than pairs of chemically similar compounds.

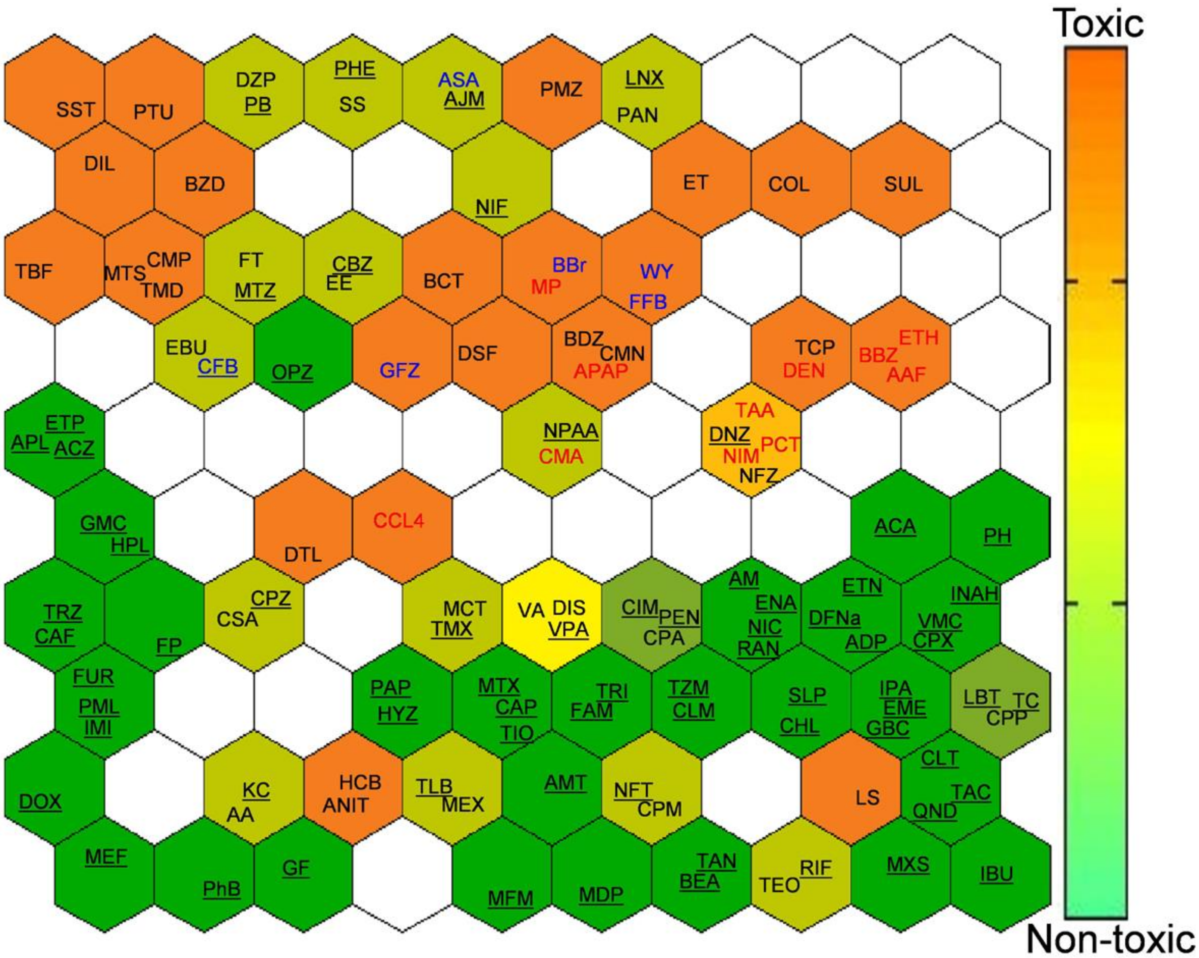


Figure 2.5. Self-organizing map of the compounds clustered by the expression of the 85 selected transcripts. Nontoxic (underlined) compounds are tightly clustered in the bottom right. PPARα activating and oxidative stress-inducing chemicals are colored in blue and red, respectively.

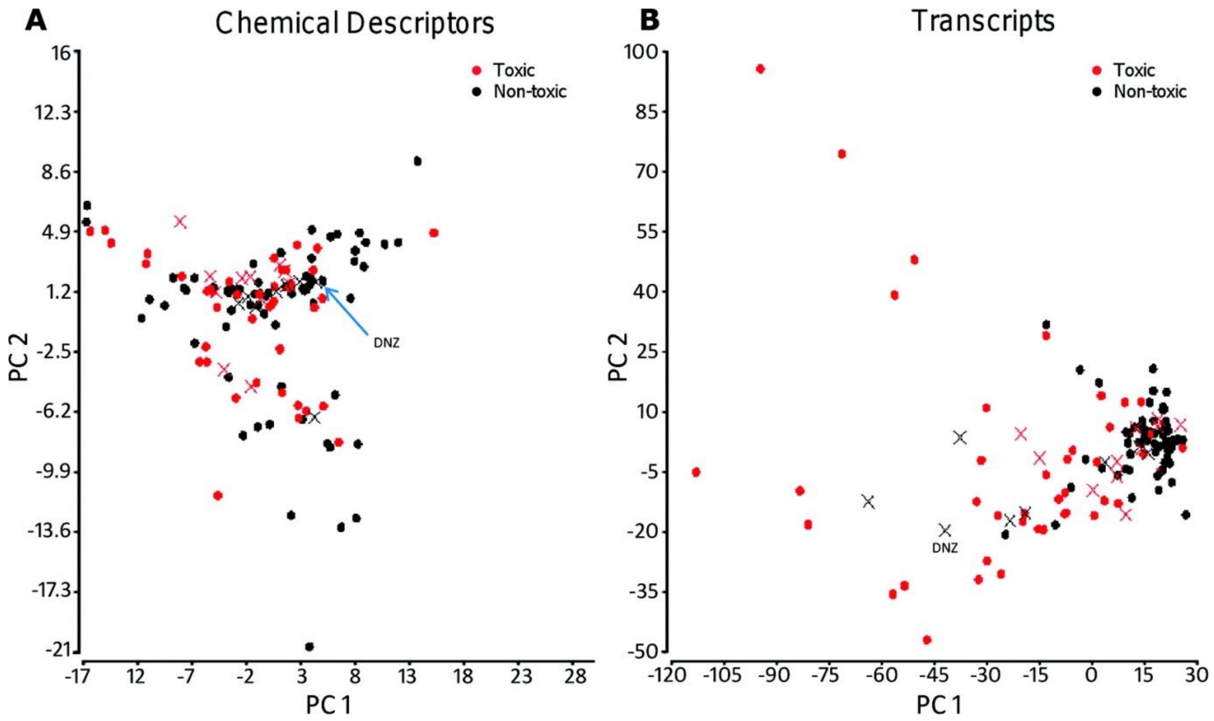
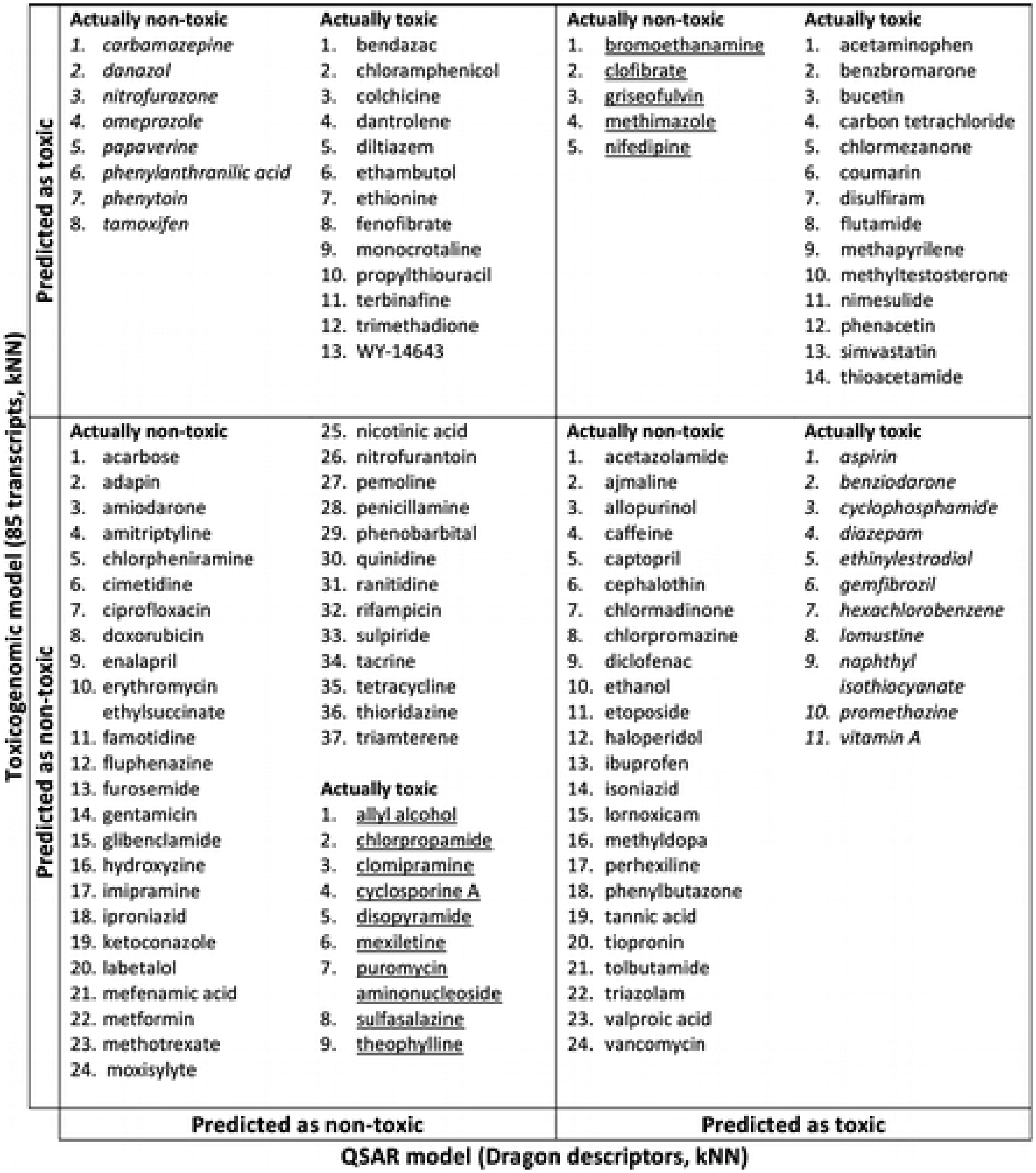


Figure 2.6. Principal component analysis of the chemical (A) and toxicogenomics (B) descriptors. Toxic and nontoxic compounds are colored red and black, respectively. Compounds mis-predicted by the toxicogenomics model but correctly predicted by the QSAR model are marked as crosses (×). An example of a nontoxic compound (danazol, DNZ) which has distant toxic toxicogenomic neighbors but close nontoxic chemical neighbors is shown.

Table 2.4. Confusion Matrix Showing Predictions by the QSAR Model and Toxicogenomics Model. Compounds mis-predicted by the toxicogenomics model but correctly predicted by the QSAR model are identified in italicized font. Compounds mis-predicted by both the QSAR model and by the toxicogenomics model are underlined.



Toxicogenomic model (85 transcripts, kNN)	Predicted as toxic	**Actually non-toxic** 1. *carbamazepine* 2. *danazol* 3. *nitrofurazone* 4. *omeprazole* 5. *papaverine* 6. *phenylanthranilic acid* 7. *phenytoin* 8. *tamoxifen* **Actually toxic** 1. bendazac 2. chloramphenicol 3. colchicine 4. dantrolene 5. diltiazem 6. ethambutol 7. ethionine 8. fenofibrate 9. monocrotaline 10. propylthiouracil 11. terbinafine 12. trimethadione 13. WY-14643	**Actually non-toxic** 1. bromoethanamine 2. clofibrate 3. griseofulvin 4. methimazole 5. nifedipine **Actually toxic** 1. acetaminophen 2. benzbromarone 3. bucetin 4. carbon tetrachloride 5. chlormezanone 6. coumarin 7. disulfiram 8. flutamide 9. methapyrilene 10. methyltestosterone 11. nimesulide 12. phenacetin 13. simvastatin 14. thioacetamide
	Predicted as non-toxic	**Actually non-toxic** 1. acarbose 2. adapin 3. amiodarone 4. amitriptyline 5. chlorpheniramine 6. cimetidine 7. ciprofloxacin 8. doxorubicin 9. enalapril 10. erythromycin ethylsuccinate 11. famotidine 12. fluphenazine 13. furosemide 14. gentamicin 15. glibenclamide 16. hydroxyzine 17. imipramine 18. iproniazid 19. ketoconazole 20. labetalol 21. mefenamic acid 22. metformin 23. methotrexate 24. moxisylyte 25. nicotinic acid 26. nitrofurantoin 27. pemoline 28. penicillamine 29. phenobarbital 30. quinidine 31. ranitidine 32. rifampicin 33. sulpiride 34. tacrine 35. tetracycline 36. thioridazine 37. triamterene **Actually toxic** 1. allyl alcohol 2. chlorpropamide 3. clomipramine 4. cyclosporine A 5. disopyramide 6. mexiletine 7. puromycin aminonucleoside 8. sulfasalazine 9. theophylline	**Actually non-toxic** 1. acetazolamide 2. ajmaline 3. allopurinol 4. caffeine 5. captopril 6. cephalothin 7. chlormadinone 8. chlorpromazine 9. diclofenac 10. ethanol 11. etoposide 12. haloperidol 13. ibuprofen 14. isoniazid 15. lornoxicam 16. methyldopa 17. perhexiline 18. phenylbutazone 19. tannic acid 20. tiopronin 21. tolbutamide 22. triazolam 23. valproic acid 24. vancomycin **Actually toxic** 1. *aspirin* 2. *benziodarone* 3. *cyclophosphamide* 4. *diazepam* 5. *ethinylestradiol* 6. *gemfibrozil* 7. *hexachlorobenzene* 8. *lomustine* 9. *naphthyl isothiocyanate* 10. *promethazine* 11. *vitamin A*
		Predicted as non-toxic	**Predicted as toxic**
		QSAR model (Dragon descriptors, kNN)	

The best hybrid model had similar performance to the best toxicogenomics model (76–77% CCR), differing only in the predictions of three compounds (ajmaline, griseofulvin, propylthiouracil). Examining QSAR and toxicogenomics models in comparison with each other revealed instances for which the models were complementary. When both QSAR and toxicogenomics models were in agreement, it implied greater reliability of the prediction (Table 2.4). When predictions made with these two types of models were in disagreement, deferring to the toxicogenomics model (statistically superior to the QSAR model) would more likely return correct predictions. However, of note were 19 compounds (italicized in Table 2.4) mis-predicted by the toxicogenomics model but correctly predicted by the QSAR model. The PCA plot shows that many of these compounds (denoted by crosses in Figures 2.6A and 2.6B) had neighbors in the multidimensional toxicogenomics descriptor space of opposite toxicities (Figure 2.6B), but their neighbors in the chemistry space had similar toxicities (Figure 2.6A). For example, nontoxic danazol has toxic neighbors in the toxicogenomics descriptor space (Figure 2.6B) but nontoxic neighbors in the chemistry space (Figure 2.6A). Some of these mis-predicted compounds, *e.g.*, gemfibrozil (PPARα activator) and lomustine (genotoxic hepatocarcinogen), exhibit late-onset toxicity which could explain the failure of 24 h expression profiles to capture relevant changes and consequently to predict their 28-day hepatotoxicity.

2.5. Discussion

Our study showed that chemical features and toxicogenomics data were useful and relevant for the development of classification models for understanding and predicting hepatotoxicity. The high classification accuracy of toxicogenomics models supports the use of early transcriptional response as an indicator for long-term toxicity and for understanding a potential mode of action. Even though QSAR models were less predictive, they will continue to be used for initial virtual screening in cases where no experimental data (*e.g.*, toxicogenomics) are available. By developing hybrid models using both chemical descriptors and toxicogenomics data, we identified both chemical features and transcripts, which provided additional insights into understanding drug-induced liver injury.

2.5.1. Biological Pathways Involved in Liver Injury

Toxicogenomics data from single exposure were not only useful for the classification of 28-day liver injury phenotype but also provided important mechanistic insights into pathways that may lead to long-term toxicity. Pathway analysis showed that the 85 most predictive transcripts were in *Hnf4α*-, *Myc*-, and *Eif2*-centered networks, all of which have been implicated in hepatotoxicity. *Hnf4α*, a transcriptional factor of the nuclear hormone receptor family, is known to play an important role in liver function, morphological and functional differentiation of hepatocytes, cell proliferation, and detoxification (Parviz et al. 2003). Although the *Hnf4α* gene itself was not among the selected transcripts, *Hnf4α*-regulated genes were up-regulated in the early stage of hepatocellular injury.

In addition, *Hnf4α* is essential for controlling the acute phase response of the liver induced by endoplasmic reticulum (ER) stress (Luebke-Wheeler et al. 2008). ER stress is a common response to many toxicants, and under conditions of severe or prolonged ER stress, apoptosis is triggered by accumulation of incompletely assembled or misfolded proteins (Ji & Kaplowitz 2006). Activation of Eif2 signaling pathway is widely recognized as a key contributor to ER stress. In the present study, we found the characteristic up-regulation of several genes involved in Eif2 signaling pathway after treatment with several hepatotoxicants, such as Eif2 subunit 1 alpha (*Eif2s1*), Eif3 subunits G (*Eif3G*) and J (*Eif3J*), and *Eif4a1*. Thus, our analysis provided additional supporting evidence that the *Eif2* signaling pathway may be a common mechanism involved in early liver damage through ER stress.

Myc is a transcription factor which regulates cell proliferation, differentiation, and apoptosis (Lin et al. 2009). In the present study, we found up-regulations of several genes in the Myc-centered network including transcription factors nucleophosmin 1 (*Npm1*), TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor (*Taf9*), *Eif4a1*, and general transcription factor IIIC polypeptide 3 (*Gtf3c3*). While further studies are needed to link the effects of individual chemicals to transcriptional changes in the Myc-centered network, our analysis shows that these transcripts may be important early predictive biomarkers for subchronic hepatocellular injury.

Biological pathway analysis revealed the down-regulation of genes involved in cellular function and maintenance, consisting of transporters and inflammatory response, such as the complement system pathway. Abnormal homeostasis and cellular function are often associated with hepatotoxicity. In particular, coagulopathy is often involved because many factors in the coagulation system are synthesized in the liver. Recently, toxicogenomics biomarkers for diagnosis and prognosis of hepatotoxicity-related coagulation abnormalities have been reported (Hirode et al. 2009b). Our results further support that malfunction of the coagulation system is a common feature in liver injury and that the down-regulation of complement 8, β-polypeptide (*C8b*), complement 9 (*C9*), and complement factor B (*Cfb*) may be an early indicator of impaired liver function by different types of drugs.

Many of the 85 selected transcripts have also been previously implicated with liver diseases by the same chemicals in the Comparative Toxicogenomics Database (http://ctd.mdibl.org/). For instance, ubiquitin specific peptidase 10 (*Usp10*) has been associated with the Myc-centered network in acetaminophen-induced liver toxicity (Beyer et al. 2007). It is also closely related to ubiquitin specific peptidase 2 (*Usp2*) which is among the 37 genes used to derive a toxicogenomics model for hepatotumorigenesis by (Fielden et al. 2007). The agreement with previous findings lends credence to our selected list of transcripts as biomarkers for hepatotoxicity.

2.5.2. Hybrid Models Afford More Reliable Exploration of Chemical Structural Alerts

Development of QSAR models of hepatotoxicity for structurally diverse chemicals is a challenge (Rodgers et al. 2010), and the results of this study show that a correct classification rate of such models ranged between 55 and 61%. Thus, interpretation of such models with regards to the potential chemical "structural alerts" for hepatotoxicity may be futile. However, when chemical descriptors and toxicogenomics data were used together to develop hybrid models, significantly higher predictive accuracy (as high as 77%) of the models provided additional confidence for considering the chemical fragments selected by the models as potentially predictive of an increased risk of liver toxicity. By examining the chemical substructures suggested by the hybrid models (see Table 2.3), we observe that features selected through the modeling procedure are several well-known toxicophores. This finding provides a strong indication of the value of hybrid modeling for identification of the toxicophores as compared to the traditional QSAR, which is plagued by a weaker predictive power.

2.5.2.1. Substructure A (Acetanilide): Toxic Species Formed, N-Hydroxylamines and Nitroso Compounds

The acetanilide substructure was present in several hepatotoxic drugs, as well as the nontoxic phenylbutazone. The acetanilide substructure is especially susceptible to N-oxidation (Loew & Goldblum 1985). The N-hydroxylamine and nitroso products are highly reactive. However, some compounds may be toxic due to activation at sites outside of the acetanilide substructure. For example, acetaminophen owes much of its toxicity to the quinone imine metabolite despite its chemical similarity with phenacetin. Its only difference from phenacetin is its 4-hydroxyl group, which is preferentially oxidized by CYP2E1 to the reactive quinone imine. In phenacetin and bucetin, the 4-hydroxyl group is replaced by an alkoxyl substituent which renders them less susceptible to quinone formation and more likely to be activated by N-hydroxylation (Peters et al. 1999). Phenylbutazone also undergoes another transformation (aromatic hydroxylation) instead of N-hydroxylation (Aarbakke et al. 1977). This probably explains its lack of rat hepatotoxicity in this study despite containing the acetanilide substructure.

2.5.2.2. Substructure B (Thioamide): Toxic Species Formed, Sulfur Species of Various Oxidation States

Our models showed that the presence of thioamide (Table 2.3, substructure B) is associated with hepatotoxicity. Thiocarbonyls are often oxidized or desulfurated to produce toxic sulfur-containing species. Thioacetamide S-oxide is highly polar and forms adducts with proteins (Porter & Neal 1978). Disulfiram, despite being a dithiocarbamate instead of a thioamide, also forms a sulfoxide that binds to proteins and inhibits their activity. Such protein binding is also responsible for disulfiram’s therapeutic inhibition of aldehyde dehydrogenase (Shen et al. 2001). The only nontoxic drug that has this substructure was methimazole. Although methimazole was defined as nonhepatotoxic in this study, it has been reported to yield atomic sulfur species that bind and inhibit P450 activity, possibly leading to liver necrosis (Lee & Neal 1978).

2.5.2.3. Substructure C (Alkyl Chloride): Toxic Species Formed, Alkyl Radicals

Hepatotoxicity of alkyl chloride compounds has been attributed to the homolytic cleavage of the C–Cl bond which produces damaging free radicals, especially among highly halogenated compounds. This is a well-studied phenomenon best exemplified by carbon tetrachloride and its alkyl halide analogs such as chloroform and bromotrichloromethane (Rechnagel & Glende 1973). However, other chlorinated alkanes studied here, cyclophosphamide, lomustine and chloramphenicol, do not share the same toxic mechanism as carbon tetrachloride and cannot be attributed to the C–Cl bond. For instance, the ultimate toxicant responsible for cyclophosphamide hepatotoxicity is acrolein, which is formed independently of the alkyl chloride group.

2.5.2.4. Substructure D (Styrene): Toxic Species Formed: Epoxides

The nonaryl double bond in substructure D when it is part of a benzofuran or benzopyran is especially prone to epoxide formation (Kaufmann et al. 2005). Such epoxides often form DNA and protein adducts (Adam et al. 1993). Coumarin’s toxicity requires the formation of an epoxide, which is followed by subsequent rearrangement of the epoxide to o-hydroxyphenylacetaldehyde, which is considered to be the hepatotoxic intermediate (Vassallo et al. 2004). Hence, it is comparatively more toxic in rats than in humans because of the rat’s metabolism via the 3,4-epoxide (Lake et al. 1989), while in humans, coumarin primarily undergoes aromatic hydroxylation instead of forming the above-mentioned epoxide (Felter et al. 2006, Vassallo et al. 2004). The three benzofurans in our study, benziodarone, benzbromarone, and amiodarone, are known hepatotoxic agents whose toxicity has been attributed to the 2-substituted benzofuran (Kaufmann et al. 2005). Although amiodarone was not found to be hepatotoxic on the basis of its 28-day histopathology and serum chemistry results, hepatocellular vacuolization indicative of phospholipidosis was noted (Appendix 1 Table A1.2.1).

2.5.3. Limitations

The performance of QSAR models generally suffers when predicting complex toxicity end points such as hepatotoxicity, a phenotype with several complex mechanisms. There are numerous examples of chemically similar compounds with widely divergent liver effects. While ibuprofen is safe in humans, ibufenac, lacking a methyl group, is toxic (Rodgers et al. 2010). In our data set, nontoxic caffeine and toxic theophylline differ by a methyl group. This phenomenon is known as an "activity cliff" where very similar molecules possess disparate activities, such that the profile of activity plotted against compound's similarity is akin to a rugged landscape with many cliffs (Maggiora 2006). QSAR can be realistically applied if there are enough compounds to adequately represent the complex activity landscape. Unfortunately, this was not the case for our data set. The high proportion (50%) of opposite activities among chemically similar pairs compounded by the lack of congeners in our chemically diverse set posed further challenges to QSAR modeling. Hence, it was not surprising that the CCR of the QSAR models could barely exceed 60% in predicting the biologically complex hepatotoxicity end point.

2.6. Conclusions

In conclusion, this study shows that while QSAR and toxicogenomics are both important predictive tools on their own, concomitant exploration in chemical and toxicogenomics descriptor spaces, through hybrid models, will elicit deeper insight. Consistent with results from other toxicogenomics studies, we showed that toxicogenomics is predictive and provides valuable mechanistic information. The pathways suggested several mechanisms such as ER stress and coagulopathy that could be related to hepatotoxicity. As QSAR is entirely computational and obviates the need for experiments, it will remain an important virtual screening tool. Importantly, structural alerts can be identified with greater confidence from the better fitted hybrid models. In addition, hybrid models improve and refine the interpretation of the data in terms of chemical alerts for hepatotoxicity. Additional studies using methodologies and descriptors that can handle activity cliffs in both chemical and toxicogenomics descriptor spaces may improve the predictive power of models developed in this study and exploit further the complementarities between QSAR and toxicogenomics models of hepatotoxicity.

CHAPTER 3. INTEGRATIVE CHEMICAL-BIOLOGICAL MODELING WITH NEW METHODS: INTEGRATIVE CHEMICAL AND BIOLOGICAL READ-ACROSS (CBRA) FOR TOXICITY PREDICTION[2]

3.1. Overview

Traditional read-across approaches typically rely on the chemical similarity principle to predict chemical toxicity; however, the accuracy of such predictions is often inadequate due to the underlying complex mechanisms of toxicity. Here we report on the development of a hazard classification and visualization method that draws upon both chemical structural similarity and comparisons of biological responses to chemicals measured in multiple short-term assays (”biological” similarity). The Chemical-Biological Read-Across (CBRA) approach infers each compound’s toxicity from those of both chemical and biological analogs whose similarities are determined by the Tanimoto coefficient. Classification accuracy of CBRA was compared to that of classical RA and other methods using chemical descriptors alone, or in combination with biological data. Different types of adverse effects (hepatotoxicity, hepatocarcinogenicity, mutagenicity, and acute lethality) were classified using several biological data types (gene expression profiling and cytotoxicity screening). CBRA-based hazard classification exhibited consistently high external classification accuracy and applicability to diverse chemicals. Transparency of the CBRA approach is aided by the use of radial plots that show the relative contribution of analogous chemical and biological neighbors. Identification of both chemical and biological features that give rise to

[2] This chapter is a reproduction, in whole, with permission of a publication in Chemical Research in Toxicology by (Low et al. 2013a) at doi:10.1021/tx400110f.

the high accuracy of CBRA-based toxicity prediction facilitates mechanistic interpretation of the models.

3.2. Introduction

The chemical similarity principle (Johnson & Maggiora 1990, Willett et al. 1998) posits that chemically similar compounds are likely to exhibit similar effects. Consequently, a variety of chemical similarity-based methods have been developed to predict chemical-induced responses from chemical structures alone. The chemical similarity principle provides the basis for both straightforward read-across analysis (Enoch et al. 2008, Hewitt et al. 2010, Schüürmann et al. 2011, Wang et al. 2012, Wu et al. 2010), and more complex machine learning-based approaches used in Quantitative Structure-Activity Relationship (QSAR) modeling (Gleeson et al. 2012, Voutchkova et al. 2010, Zvinavashe et al. 2008).

Chemical structure-based prediction methods face limitations, especially when the challenge is to accurately predict complex *in vivo* outcomes (Gleeson et al. 2012, Nikolova & Jaworska 2003). Data from *in vitro* screening of thousands of chemicals in hundreds of experimental systems provide additional biological activity information at molecular and cellular levels potentially useful for predictive toxicology modeling (Judson et al. 2012, Rusyn et al. 2012, Valerio & Choudhuri 2012). Indeed, integration of chemical structural features and biological screening data provides important advantages over traditional QSAR modeling, such as improved prediction accuracy (Sedykh et al. 2011, Zhu et al. 2008), greater coverage of chemical space and a better interpretation of chemical and biological features (Low et al. 2011).

While QSAR modeling approaches have grown in popularity and complexity, end-users often show preference for simple and more transparent methods such as read-across, *e.g.*, OECD QSAR Toolbox (http://www.qsartoolbox.org/). The read-across methodology requires chemical (*i.e.*, structure-based) similarity as a starting point. The objective of this method is to predict the toxicity behavior of a compound (*i.e.*, produce an equivalent of a test result) by inferring from structurally similar chemicals with available toxicity data. Grouping and read-across of chemicals in a hazard and/or risk assessment context are well established and can be used to satisfy information requirements under Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation in the European Union. For example, more than 20% of high production volume chemicals submitted for the first REACH deadline relied on read-across for hazard information on a number of toxicity endpoints necessary for registration (http://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2011_en.pdf). Under REACH, flexibility exists for how the analogs are selected; however, the read-across argument needs to be convincingly substantiated with scientifically credible justification. Thus, even though chemical structure-based read-across represents an alternative to standard animal-based tests, the inherited uncertainty of the prediction, combined with a lack of a standardized framework for its application by the decision-makers, creates a need to increase confidence in prediction and utilize visual aids for presenting evidence in a transparent manner.

Although chemical and biological factors are sometimes considered using a weight-of-evidence framework (Hewitt et al. 2010, Wang et al. 2012, Wu et al. 2010), these approaches are largely qualitative, not completely transparent, and may be prone to bias. Hence, there is a need to automate a read-across process that would combine both chemical and biological factors and yet, keep the process transparent for expert interrogation. To that end, we introduce a quantitative toxicity prediction approach combined with a visualization methodology, termed chemical-biological read-across (CBRA), that relies not only on inherent chemical properties (*chemical descriptors*), but also on biological profiles measured by short-term experimental assays (*biological descriptors*). A graphical display of the compound's classification, along with the identity of the neighbors and weights applied, is employed to increase transparency and interpretability. Using several data sets with short-term bioassay profiles, we demonstrate the advantages of CBRA over other methods that rely on biological and/or chemical descriptors alone.

3.3. Materials and methods

3.3.1.Data sets

Four data sets were used in this study (descriptor matrices and prediction endpoints of compounds are available as supplemental material of the online publication). The first data set contained 127 compounds from the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system (TG-GATES). The target property for prediction is sub-chronic hepatotoxicity previously modeled in (Low et al. 2011) based on liver histopathology and clinical chemistry findings over 28 days of repeat dosing (Uehara et al. 2010). Gene expression data (biological features) and chemical descriptors (inherent chemical features) were processed as explained in (Low et al. 2011). Briefly, of the 31,042 probes on the arrays, we removed those that were consistently not expressed or did not change their expression values across all compounds between treated vs. vehicle control groups. Next, 2,991 transcripts were selected that varied in their expression across all the compounds based on the following criteria: the largest change of any transcript over its untreated equivalent was over 1.5 fold and the smallest false discovery rate (Welch t-test) was less than 0.05. Then, transcripts with low variance (all, or all but one value is constant) were excluded and further, one of each pair of transcripts with high pairwise correlation (r^2>0.9) chosen randomly, was removed; this left 2,923 transcript variables which were range scaled and used for model building. Further, genes differentially expressed between hepatotoxic and nonepatotoxic compounds were ranked by a permutation t-test (signfiicance analysis of microarrays). To avoid feature selection bias, such supervised feature selection was performed according to the same 5-fold external cross-validation scheme later used for model evaluation. In the end, we selected 85 unique transcripts (5 sets of 30 top-ranked transcripts per fold).

The second data set contained 132 compounds (DrugMatrix®, https://ntp.niehs.nih.gov/drugmatrix). The biological descriptors were the expression of 200 genes in the rat liver (5 day repeat dosing), that were selected as detailed in (Natsoulis et al. 2008). The data for the prediction target, hepatocarcinogenicity, was compiled in a related study by (Fielden et al. 2007) using literature sources and the Carcinogenicity Potency Database (CPDB, http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html).

The third and fourth data sets were from (Lock et al. 2012) in which 240 compounds were tested for cytotoxicity (intracellular ATP and caspase-3/7 apoptosis) in 84 lymphoblastoid cell lines with different genotypes. The 148 biological descriptors used here were the consolidated cytotoxicity profiles derived as detailed in (Sedykh et al. 2011). Of the original 240 compounds, 185 compounds were assigned mutagenicity labels according to CCRIS (http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS) and 122 compounds were assigned rat oral LD_{50} according to ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus/). Because CCRIS provides detailed-level mutagenicity test results (including test strains, concentrations, and type of metabolic activation), the final mutagenicity assignments were determined using the protocol described in (Mortelmans & Zeiger 2000): mutagens must test positively in any one of the five standard Ames *Salmonella* test strains; non-mutagens, in contrast, must be consistently negative (in at least 4 out of 5 test strains). Additionally, when Ames *Salmonella* strains beyond the standard five were available, we defined mutagens as those tested positively in at least 20% of the strains and non-mutagens as those tested negatively in at least 80% of the strains regardless of metabolic activation. For the other classification endpoint, continuous rat oral LD_{50} values were split into two classes based on a threshold of 300 mg/kg, consistent with the threshold separating categories 1-3 and 4-6 in the Globally Harmonized System of Classification and Labeling of Chemicals (UNECE 2009).

3.3.2. Chemical descriptors and data processing

Chemicals utilized in all data sets underwent structural curation according to the procedures described in (Fourches et al. 2010). This involved standardizing the molecular structures and removing salts, duplicates and problematic structures (*e.g.*, metal-containing, molecular weight > 2000). Next, Dragon (v.5.5, Talete SRL, Milan, Italy) descriptors were computed for all chemicals.

After additional treatment of gene expression data (data sets 1 and 2, see above), all chemical and biological descriptors were range scaled to fall between 0 and 1. Furthermore, descriptors with low-variance (standard deviation $<10^{-6}$) or one of any pair of descriptors with high intercorrelation (pairwise $r^2 > 0.9$) were removed.

3.3.3. Quantitative read-across methodology

For read-across, the predicted activity of a compound (A_{pred}) was calculated using the following equation (Equation 1) from the similarity (S_i) weighted aggregate of the activities A_i of k nearest neighbors.

$$A_{pred,RA} = \frac{\sum_{i=1}^{k} S_i \cdot A_i}{\sum_{i=1}^{k} S_i} \qquad 1]$$

The pairwise Tanimoto similarity, S_i, between the molecule of interest (A) and its *i*th neighbor (B), was calculated from the Jaccard distance d_{Jac} [Equation 2, (Willett et al., 1998)] across descriptors $x_1, \ldots, x_p$. For a set of range-scaled continuous descriptors, Tanimoto similarity is normalized between 0 and 1 with 1 corresponding to identical pairs.

$$S_i = 1 - d_{Jac} = \frac{\sum_{m=1}^{p} x_{A,m} \cdot x_{B,m}}{\sum_{m=1}^{p} {x_{A,m}}^2 + \sum_{m=1}^{p} {x_{B,m}}^2 - \sum_{m=1}^{p} x_{A,m} \cdot x_{B,m}} \qquad 2]$$

The similarity-weighted aggregate in Equation 1 ensures that the activities of more similar neighbors are given higher weights when calculating the predicted activity.

For CBRA, compound activity was estimated from sets of neighbors in both the biological (*bio*) and chemical (*chem*) data (Equation 3).

$$A_{pred,CBRA} = \frac{\sum_{i=1}^{k_{bio}} S_i \cdot A_i + \sum_{j=1}^{k_{chem}} S_j \cdot A_j}{\sum_{i=1}^{k_{bio}} S_i + \sum_{j=1}^{k_{chem}} S_j} \qquad 3]$$

Two sets of nearest neighbors (k_{bio} biological neighbors in the biological space characterized by bioassay profiles and k_{chem} neighbors in the chemical space characterized by Dragon descriptors) were used for the estimation of the toxicity for each test compound. Activities of nontoxic compounds were assigned “-1” while those of toxic compounds were assigned “+1”. The predicted classification threshold was set at zero such that compounds with negative predicted activity were considered as nontoxic and toxic otherwise.

Read-across was performed in two ways depending on how the maximum number of neighbors was determined: 1) by a similarity threshold (RA-sim), or 2) by a set value of *k* (RA-*k*NN). RA-sim included all neighbors with similarity greater than or equal to a similarity threshold set at 0, 0.6, 0.7, 0.8, or 0.9; RA-*k*NN included only *k* nearest neighbors (possible *k* values: integers from 1 to 5). Limiting the number of nearest neighbors by 5 is arbitrary; generally it should be understood that the selection of a large number of nearest neighbors would undermine the nearest neighbor selection principle so "5" is a threshold number we have been using typically in our implementation of the *k*NN QSAR method (Zheng & Tropsha 2000).

Other models using both biological and chemical descriptors

In addition to CBRA, other approaches combining biological and chemical data for toxicity prediction were examined. First, biological and chemical descriptors were pooled together constituting a "hybrid" space which generated a single set of *k* nearest neighbors for each molecule (see Equation 1 where $k=k_{hybrid}$). Second, compounds' activities were predicted by developing independent biological read-across and chemical read-across models and pooling the resulting predictions, essentially forming an ensemble model (Equation 4).

$$A_{pred,ensemble} = \frac{1}{2}\left(\frac{\sum_{i=1}^{k_{bio}} S_i \cdot A_i}{\sum_{i=1}^{k_{bio}} S_i} + \frac{\sum_{j=1}^{k_{chem}} S_j \cdot A_j}{\sum_{j=1}^{k_{chem}} S_j}\right) \qquad 4]$$

3.3.4. Model evaluation

External 5-fold cross validation.

All models were evaluated using an external 5-fold cross validation. Briefly, each data set was randomly divided into five equal parts with the same toxic/nontoxic ratio before modeling. Each of the five parts was left out in turn to form an external set for validating the model developed on the remaining four parts (modeling set). For each target compound in the external set, neighbors were selected from the modeling set and not from the external set.

Internal 10-fold cross validation in read across kNN method

In RA-kNN, optimal k_{bio} and k_{chem} were selected from values between 1 and 5 by additional internal 10-fold cross-validation. Briefly, each modeling set was further divided according to a 10-fold cross-validation scheme, forming ten pairs of training and test sets. For each training set, we performed a grid search across all 25 possible pairs of k_{bio} and k_{chem} values and validated the models using the corresponding test set. For each pair of k_{bio} and $\mathrm{k}_{\mathrm{chem}}$ values, the balanced accuracies across the ten test sets were averaged. The pair yielding the highest mean balanced accuracy was considered to be optimal and its k_{bio} and k_{chem} values were subsequently applied to the corresponding external validation set. Therefore, the five modeling sets resulted in five optimal models with various k_{bio} and k_{chem} values.

y-randomization

The *y*-randomization test (randomization of the response) was performed to ensure that models were robust and not due to chance correlations. After random permutation of the activity labels in the modeling sets, models were rebuilt following the same workflow as described above. This protocol was repeated 30 times. Performance of the models generated from the permuted labels was compared to that of the models derived from the original data sets. Statistical significance of the difference in balanced accuracy was determined with one-tailed one-sample t-test.

Prediction performance metrics

All metrics characterizing model performance [i.e., specificity, sensitivity, accuracy, balanced accuracy, and area under curve (AUC)] were obtained from external 5-fold cross-validation. Metric values close to 1 indicate high classification accuracy while 0.5 serves as the random baseline for binary classification. Specificity is the fraction of compounds predicted correctly within the nontoxic class; conversely, sensitivity is the fraction of compounds predicted correctly within the toxic class. Accuracy is the fraction of compounds predicted correctly in total. Balanced accuracy is the average of the rates correctly predicted within each class ((specificity + sensitivity)/2). AUC is the area under the receiver operating characteristic curve of sensitivity against (1-specificity). Thus, AUC is a function of sensitivity and specificity, providing an overall accuracy metric independent of a predefined activity threshold unlike the other prediction metrics which were calculated using a predefined activity threshold of zero (for activity values ranging between -1 to +1).

Coverage of the models is reported as the fraction of compounds in the external set that are within the applicability domain (AD) for which reliable predictions are expected to be obtained. In RA-sim, a target compound is within the AD if there exists at least one neighbor in the modeling set whose similarity is above the similarity threshold; in RA- *k*NN, a compound is within the AD if there exists at least one neighbor with a minimum similarity of 0.3. Standard errors were calculated by the bootstrap method (Efron & Tibshirani 1986) using 1000 sampling trials.

3.3.5. Model interpretation

Identification of informative descriptors

Adapted from the local importance score used in the random forest method (Breiman 2001), we use a local importance score based on *x*-randomization to rank descriptors by their contribution to a target compound's predicted activity. *x*-randomization involves the random permutation of a descriptor *x* across the modeling set such that the descriptor's effect on the model's accuracy before and after permutation is compared. This difference is expected to be more pronounced for important descriptors. Specifically, after permutation, the similarity between the target compound and its *k* neighbors (previously used for RA-*k*NN) will change. This resultant change in similarity is averaged over 99 random permutations to obtain the local importance score *I*(*x*,*compound*) which measures the descriptor *x*'s contribution towards the target compound's predicted activity. This procedure was repeated for each descriptor per target compound. A high local importance score indicates that the descriptor is highly contributory to the target compound's prediction.

Visualization of nearest neighbors using radial plots

To visualize the information used to generate the predicted activity of a compound (A_{pred}), the radial plot is used (see Figures 3.1-3 for examples). The central node marks the target compound. Surrounding it are nodes representing biological neighbors (left hand side) and chemical neighbors (right hand side), all colored according to their known toxicity assignments (red=toxic, black=nontoxic). The relative position of each neighbor from the central node (*i.e.*, edge length) reflects the Jaccard distance (Equation 2) from the target compound. The nearest neighbors (shortest edges) are placed closest to the 12 o'clock position. Each radial plot displays all the neighbors relevant to a compound's prediction (*i.e.*, k_{chem} chemical neighbors and k_{bio} biological neighbors above 0.3 Tanimoto similarity, consistent with the AD similarity threshold for RA-*k*NN). The algorithm for generating the radial plots was written in R Statistical Software (version 2.14; R Foundation for Statistical Computing, Vienna, Austria) and is available as supplemental material of the online publication.

3.4. Results

3.4.1. Visualisation of the chemical-biological read across classification

The premise of this study was to establish a transparent methodology for inferring a compound's potential toxicity from its biological and chemical analogs. Here, we use graphical means to illustrate how CBRA integrates information from both biological and chemical analogs of a compound to predict its toxicity. Because the relevant information (the analogs, their similarities and known toxicity assignments) can be communicated using the radial plot, CBRA offers a highly transparent and interpretive method for hazard classification.

Figures 3.1-3.3 show the radial plots of three case study compounds, classifying them as hepatotoxic or not (see Methods for description of hepatotoxicity class designation) using both their biological (similar toxicogenomic profiles) and chemical (similar structures) analogs in the TG-GATES data set. Figure 3.1 depicts the basis for classifying chloramphenicol as "toxic". The central node was colored red to denote chloramphenicol's known toxicity. On the left hand side, all five biological neighbors were labeled as toxic (red) and they are highly similar to chloramphenicol (similarities: 0.826-0.857). On the right hand side, the five closest chemical neighbors are nontoxic (black, similarities: 0.645-0.667). All neighbors' activities are aggregated according to their similarity weights by CBRA (Equation

3), yielding Apred=+0.126, *i.e.*, a “toxic” prediction concordant with the known “toxic” assignment. Figure 3.2 shows the opposite case in which the correct classification of carbamazepine (known to be nontoxic; Apred=-0.099) was due to its greater similarity with its chemical neighbors (similarities: 0.721-0.813). Figure 3.3 shows that benzbromarone’s biological and chemical neighbors were mostly toxic (red), yielding concordant predictions (Apred=+0.688), in agreement with its known toxicity.

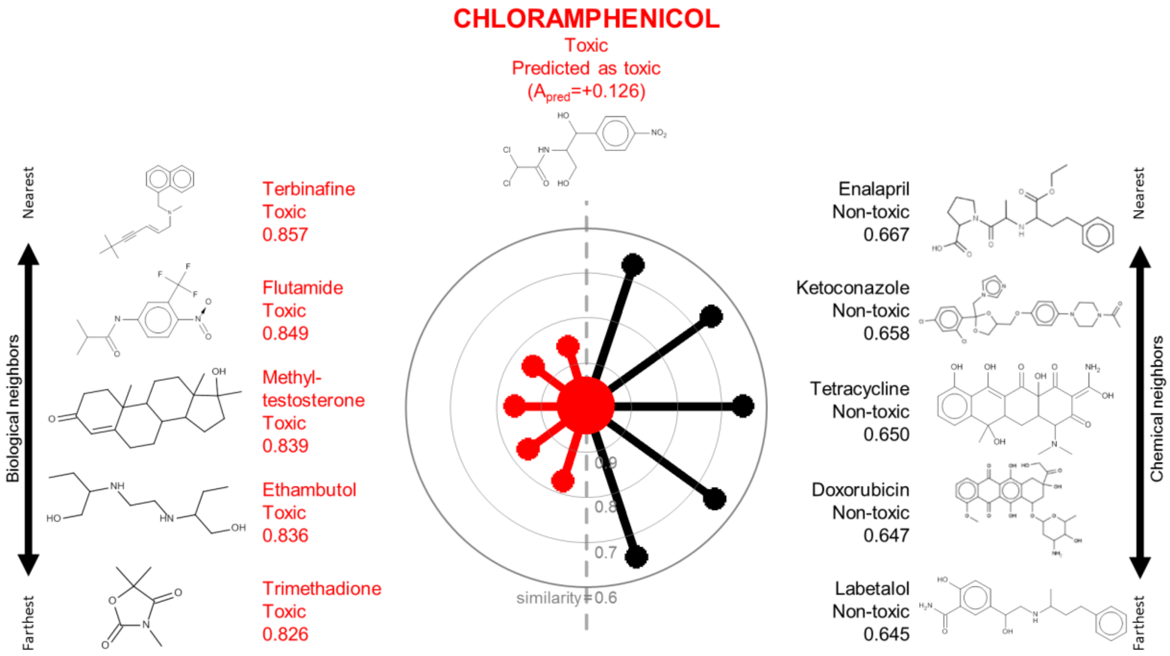


Figure 3.1. A radial plot for chloramphenicol in the TG-GATES data set.
The central node representing the target compound chloramphenicol is surrounded by biological neighbors (left hand side) and chemical neighbors (right hand side). Nearest (*i.e.*, most similar) neighbors are placed at the top. Neighbors are positioned from the target compound at a distance proportional to the Jaccard distance. Edges and nodes are colored according to the known activity classification (black: nontoxic; red: toxic).

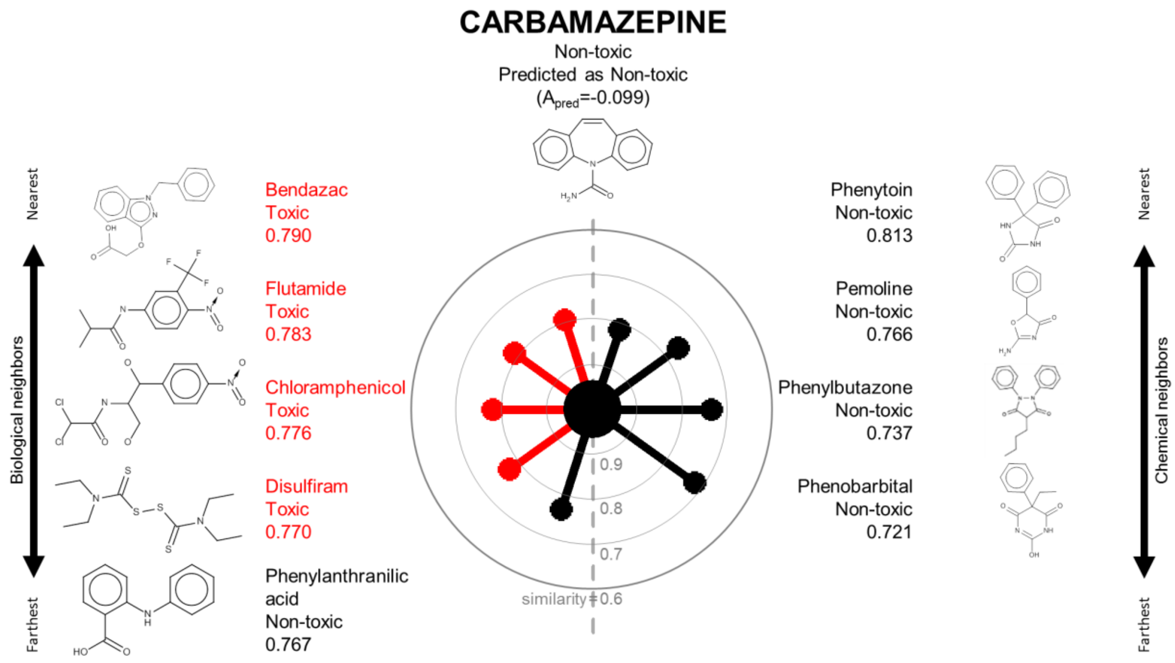


Figure 3.2. A radial plot for carbamazepine in the TG-GATES data set. See legend to Figure 3.1 for details.

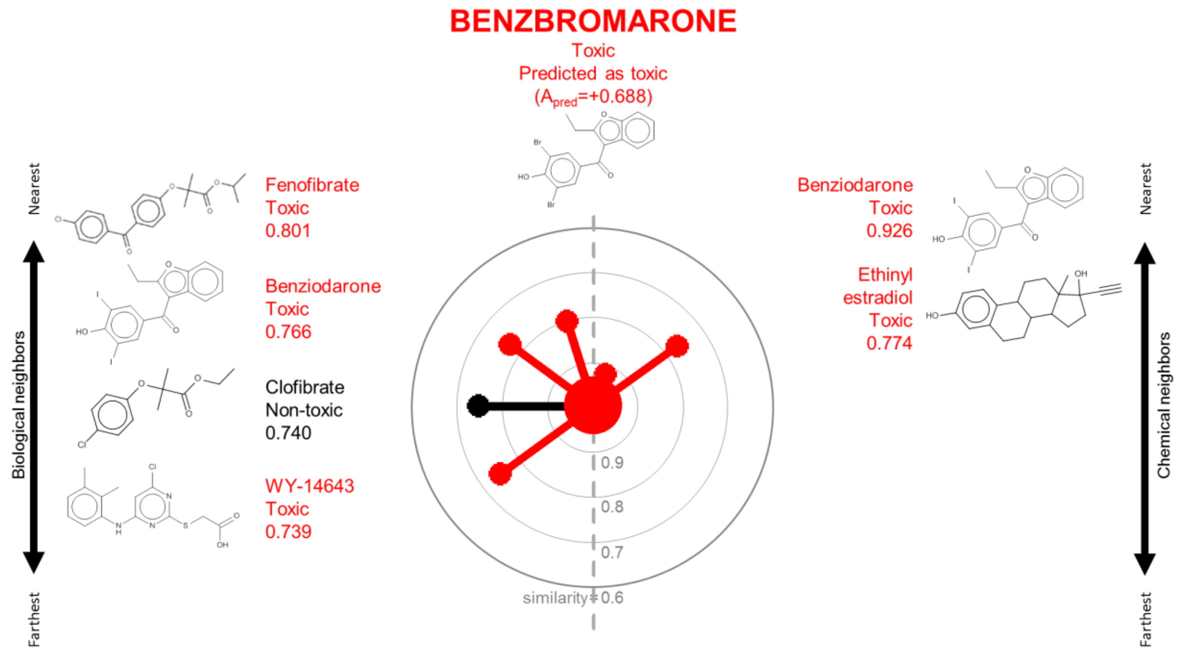


Figure 3.3. A radial plot for benzbromarone in the TG-GATES data set. See legend to Figure 3.1 for details.

Figure 3.4 provides a visual comparison of radial plots for selected compounds from the TG-GATES data set that may facilitate expert judgment of each predicted classification. As with previous radial plots, each central node represents the compound of interest and is colored according to its experiment-derived toxicity (black=nontoxic, red=toxic). Radial plots were organized by the predicted activities using only chemical neighbors (horizontal axis) and those using only biological neighbors (vertical axis). As such, the compounds can be assessed by whether the chemical or biological neighbors had higher contribution to the final classification. The lower left corner (*e.g.*, quinidine) is populated by radial plots with mostly nontoxic (black) neighbors while the upper right corner (*e.g.*, benzbromarone) is filled by those with mostly toxic (red) neighbors. Conversely, other radial plots involve discordant predictions. Such radial plots are surrounded by neighbors of various toxicities (*i.e.*, edges of various colors). Despite the discordance, the target compounds were still correctly predicted by CBRA because the activities of more similar neighbors (shorter edges) were given higher weight than those of less similar neighbors (longer edges). This simple visualization identifying neighbors based on an objective and standardized similarity metric such as the Tanimoto similarity allows users to assess the relevance of the neighbors and their contribution to the final prediction for every compound of interest.

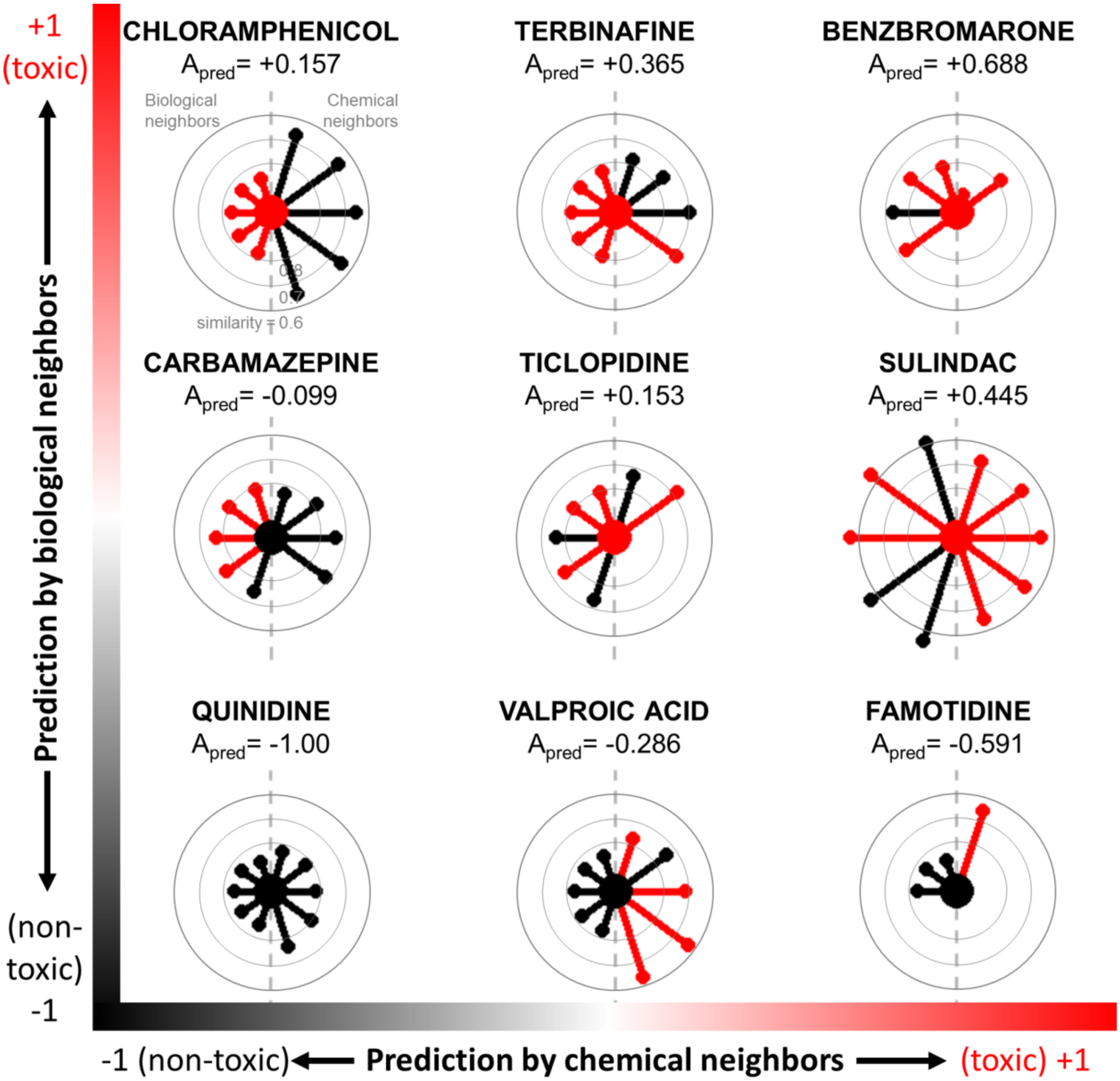


Figure 3.4. Radial plots of compounds in TG-GATES data set ordered by predictions based on chemical neighbors (horizontal axis) and biological neighbors (vertical axis). Along the horizontal axis, nontoxic predictions (black) by chemical neighbors are shown on the left and toxic predictions (red) are shown on the right. Along the vertical axis, nontoxic predictions (black) by biological neighbors are shown at the bottom while toxic predictions (red) predominate at the top. Compounds predicted with high confidence by near neighbors of similar toxicities are indicated by radial plots with short edges of the same color.

3.4.2. Model performance

<u>RA-kNN vs RA-sim</u>

As read-across can be performed in two ways using either a similarity threshold (RA-sim) or a set value of *k* (RA- *k*NN), we first compared these two approaches on the TG-GATES data set (Figure 3.5 and Appendix 1 Table A1.3.1). The first approach (RA-sim, solid filled bars in Figure 3.5) utilizing chemical descriptors only ("chemical read-across", white solid bars), showed that higher balanced accuracy may be achieved by restricting chemical similarity thresholds; however, the cost of such improved accuracy is much reduced coverage. RA-sim using gene expression data only ("biological read-across", black solid bars) had a higher balanced accuracy as compared to chemical read-across when all compounds were considered (*i.e.*, 100% coverage). However, the accuracy of biological read-across did not increase markedly when more stringent similarity thresholds were applied. Finally, CBRA (dark gray solid fill) showed the highest balanced accuracy while being the least affected by the increasing similarity threshold. RA-sim utilizing hybrid descriptors resulting from pooling both chemical and biological descriptors (light gray solid bars) together, exhibited intermediate accuracy. The second read-across approach (RA-*k*NN, patterned fill) showed comparable or higher balanced accuracy across all four types of methods integrating chemical and biological descriptors when compared to RA-sim (solid fill). For this reason, RA-*k*NN was selected as the preferred algorithm for read-across.

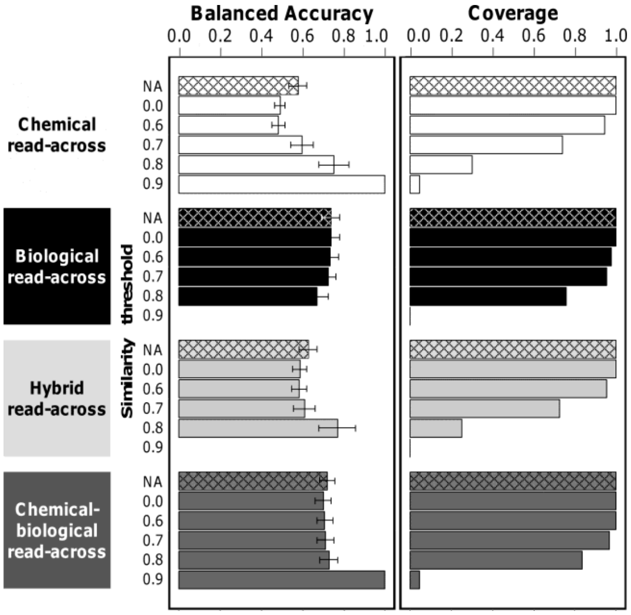


Figure 3.5. Performance of RA-sim (solid fill) and RA-*k*NN (patterned fill) models for the TG-GATES data set. RA-sim models were varied using various similarity thresholds for neighbor selection. Models based on various spaces are denoted by colors: chemical space (white), biological space (black), hybrid space (light gray), and chemical-biological spaces (dark gray).

Comparison of read-across in biological and/or chemical spaces

Next, we tested the performance of various read-across methods for different toxicity endpoints (liver toxicity and carcinogenicity, mutagenicity and acute lethality) and for different "*biological descriptor*" types (*e.g.* gene expression data from two different studies and *in vitro* cytotoxicity screening data). For this, we used both chemical and/or biological descriptors and the RA-*k*NN algorithm. In all four data sets, we applied: 1) chemical read-across (white bars); 2) biological read-across (black bars); 3) hybrid read-across by pooling chemical and biological descriptors (Equation 1, light gray bars); 4) ensemble read-across from pooling predictions from (1) and (2) (Equation 4, dark gray bars); and 5) CBRA (Equation 3, medium gray bars). Figure 3.6 and Appendix 1 Tables A1.3.1 and A1.3.2 show a comparison of the performance of the various read-across methods in each data set.

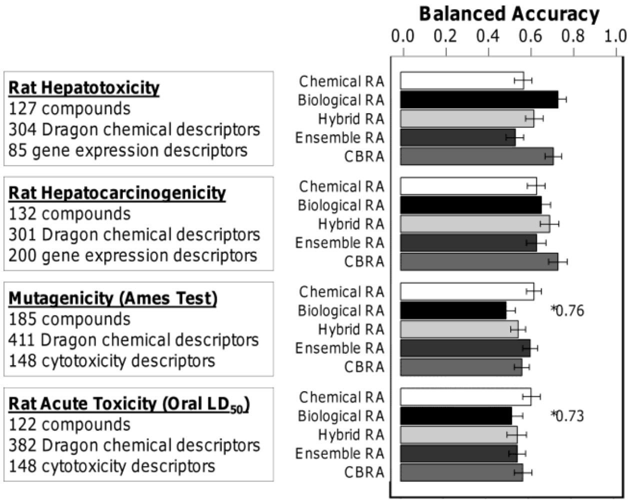


Figure 3.6. Performance of various RA-*k*NN models for four data sets. The models are colored as follows: chemical RA (white), biological RA (black), hybrid RA (from pooling chemical and biological descriptors, light gray), ensemble RA (consensus of chemical RA and biological RA, dark gray) and CBRA (gray). *Coverage values <0.95% are indicated on the chart.

While chemical read-across (white bars) exhibited highest balanced accuracy of classification for some endpoints (*i.e.*, mutagenicity and rat acute toxicity), biological read-across (black bars) had greater balanced accuracy for data sets 1 and 2 of rat hepatotoxicity and carcinogenicity, respectively, where biological descriptors represented gene expression data. However, biological read-across based on *in vitro* cytotoxicity screening data alone in data sets 3 and 4 exhibited the poorest classification accuracy (close to 50%), a result similar to that reported previously (Zhu et al. 2008). Importantly, the balanced accuracy of CBRA (medium gray bars) was consistently among the highest across all types of read-across models. Still, in three data sets (rat hepatotoxicity, mutagenicity, and rat acute toxicity), CBRA's performance, though among the best, did not surpass that of the simpler chemical read-across (white bars) or biological read-across (black bars). Similar outcomes were obtained when a comparison was made of the number of compounds correctly predicted by chemical read-across, biological read-across, and CBRA (Figure 3.7). Thus, we posit that given CBRA's consistently good performance, it should be employed where possible because it often offers the best chance of improving classification accuracy and model interpretation.

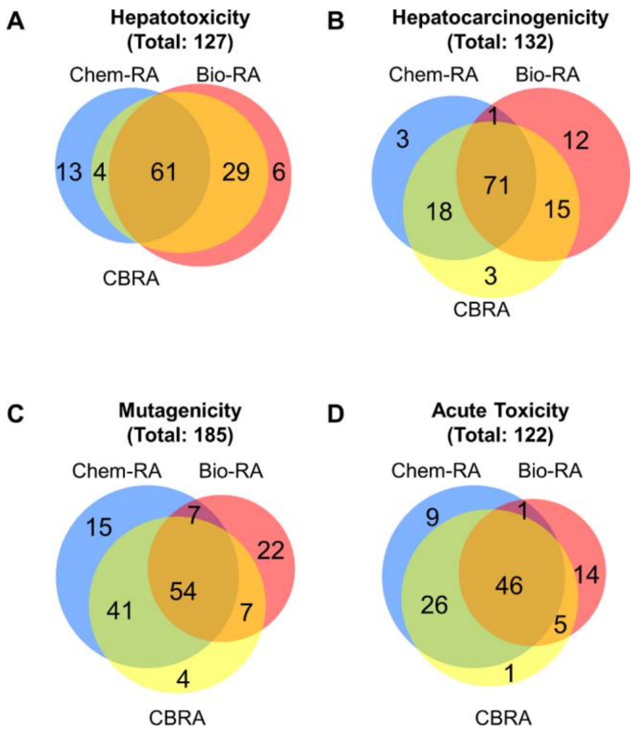


Figure 3.7. Venn diagrams depicting the number of compounds correctly predicted by chemical RA (blue circles), biological RA (red circles) and CBRA (yellow circles) in four data sets: rat hepatotoxicity (A), rat hepatocarcinogenicity (B), mutagenicity (C), and rat acute oral toxicity (D)

***y*-randomization test**

The prediction performance of all models presented in this study is given in Appendix 1 Table A1.3.1. Most models built with real data significantly outperformed those generated by *y*-randomization (p-value < 0.05) and hence, were unlikely to be fitted by chance. There were two exceptions, however, *i.e.*, models whose balanced accuracies were very poor (50%, 52%), indistinguishable from the random baseline of 50% (Appendix 1 Table A1.3.1).

3.5. Discussion

3.5.1. Improvements due to ensemble modeling and enhanced aggregation

Ensemble models have been shown to be more accurate than their constituent models[28]. The CBRA approach, effectively an ensemble model, utilizes two distinct descriptor types, *i.e.*, chemical and biological, to increase classification accuracy and uncover associations between different types of descriptors that may characterize each compound.

Our results show that simple ensemble modeling, which gives equal weights to both chemical and biological models, is insufficient to achieve high classification accuracy, as illustrated by the modest results of the simple ensemble model (dark gray, Figure 3.6). Instead, enhanced aggregation employed by CBRA ensures that the more similar neighbors have higher weights, regardless of whether they are biological or chemical neighbors. This feature of CBRA is perhaps best exemplified by the following three case studies and their radial plots (Figures 3.1-3.3) illustrating how highly similar neighbors drive the prediction outcome. These case studies were selected to represent: 1) prediction driven by biological neighbors, 2) prediction driven by chemical neighbors, and 3) concordant predictions by biological and chemical neighbors.

3.5.2. Case study: Chloramphenicol (biological space-based prediction)

Chloramphenicol (Figure 3.1 and Appendix 1 Table A1.3.3) is an anti-bacterial drug whose hepatotoxicity was linked to oxidative stress initiated by reactive metabolites (Farombi et al. 2002). Chloramphenicol increased the level of serum enzymes, as well as caused liver hypertrophy and necrosis in treated rats in the TG-GATES studies (Uehara et al. 2010). There is greater similarity in toxicogenomics profiles between chloramphenicol and its several “toxic” biological neighbors (0.826-0.856) than that with its nontoxic chemical neighbors identified using inherent chemical properties (0.645-0.667).

The gene expression profiles of chloramphenicol and its highly similar biological neighbors across 30 genes showed a consistent gene signature (Appendix 2 Figure A2.3.1A). In contrast, chloramphenicol and its chemical neighbors were characterized by relatively dissimilar descriptor profiles (Appendix 2 Figure A2.3.1B). Several genes critical to the prediction of chloramphenicol’s activity (*Abce1*, *Tomm22* and *Bmf*) are known to be implicated in mitochondrial and cell cycle regulation processes (Appendix 1 Table A1.3.4). Such deregulation is consistent with the known oxidative stress mediated hepatotoxicity of chloramphenicol. More importantly, this analysis indicates that statistically significant features elucidated by the CBRA model agree with the existing mechanistic knowledge of the compound’s toxicity. Thus, such model interpretation by CBRA may generate hypotheses about a compound’s possible mechanisms when only short-term assays are available.

3.5.3. Case study: Carbamazepine (chemical space-based prediction)

Carbamazepine is an anti-convulsant drug that acts on neuronal voltage-gated sodium channels. In the TG-GATES data set, it was classified as non-hepatotoxic in the rat. The case of carbamazepine (Figure 3.2) contrasts with that of chloramphenicol. Whereas biological RA afforded more accurate prediction than chemical RA for chloramphenicol, the opposite was found to be true for carbamazepine. Nonetheless, CBRA, in taking a similarity-weighted aggregate of the activities of k_{bio}=5 biological neighbors and k_{chem}=5 chemical neighbors, correctly predicted carbamazepine as nontoxic.

Carbamazepine and its highly similar chemical neighbors, in addition to sharing several chemical features, also exert similar pharmacological effects. Carbamazepine's nearest chemical neighbors (phenytoin, pemoline, phenylbutazone and phenobarbital) are also anti-convulsant drugs that share a tricyclic scaffold with a polar amide group in the middle (Figure 3.2). This common chemical motif is also responsible for their anti-convulsant effects. The associated pharmacophore involves an amide moiety in the middle and a side lipophilic aryl ring for interaction with the sodium channel in order to exert the drug's anti-convulsant effects (Lipkind & Fozzard 2010, Sridhar et al. 2002).

Carbamazepine's nearest biological neighbors (bendazac, flutamide, chloramphenicol, disulfiram and phenylanthranilic acid) have few obvious commonalities. Their gene expression profiles across the 30 predictor genes showed considerable heterogeneity (Appendix 2 Figure A2.3.1C). They also induce liver injury via different mechanisms and exhibit different histopathology and blood chemistry in the TG-GATES database (Uehara et al. 2010). In this instance, biological similarity determined by 24-hour gene expression may not suffice to signal 28-day liver injury.

3.5.4. Case study: Benzbromarone (concordant chemical and biological predictions)

Benzbromarone is an anti-gout agent withdrawn from the market in 2003 due to hepatotoxicity concerns (Lee et al. 2008). In the TG-GATES data set, both its biological and chemical neighbors were predictive of its hepatotoxicity. Here, we show how CBRA can provide a prediction outcome bolstered by concordant predictions as well as postulate associations between the biological and chemical neighbors for subsequent analysis (Figure 3.3).

Benzbromarone-induced hepatotoxicity is attributed to disruptions in the mitochondrial β-oxidation of fatty acids, possibly mediated by peroxisome proliferator-activated receptor-alpha (PPARα) activation (Kunishima et al. 2007). It exhibits a gene expression profile similar to those of its biological neighbors (fenofibrate, benziodarone, clofibrate and WY-14643), all known PPARα activators. Furthermore, the genes important for predicting benzbromarone's activity (*Bcs1l*, *Tomm20*, *Abce1* and *LOC100360017*, Appendix 1 Table A1.3.4), relate to mitochondrial functions, indicative of the mitochondrial-mediated hepatotoxicity observed in benzbromarone.

In this case, benzbromarone's biological and chemical neighborhoods overlap and provide concordant predictions, possibly indicative of common biological-chemical associations between the two neighborhoods. The overlapping neighbor, benziodarone, exhibits PPAR activity similar to its biological neighbors and a lipophilic, planar structure similar to its chemical neighbors. In addition, the cross-talk between estrogen and other sex hormones and PPAR-mediated signaling is well recognized (Komar 2005), which makes the association of its chemical analog, ethinyl estradiol, plausible. Hence, such cross-inference from one neighborhood to another can still provide useful clues for formulating hypotheses about biological-chemical associations, the strength of which are dependent on the extent of the overlapping neighborhoods. Furthermore, such analysis provides a novel way for the concurrent study of chemical and biological features and their underlying interactions.

3.5.5. Advantages and Limitations of CBRA

Read-across method is based on the expectation that chemically similar molecules should elicit similar biological responses. It is worth noting, however, that whichever way the chemical similarity is defined, it always has relative meaning; that is, the similarity search exercise identifies the most similar compounds in a given set of compounds, and not necessarily the most similar chemically feasible structures. Further, structure-activity relationship landscapes are known to be "rough," with many molecules appearing chemically similar but nevertheless having rather different biological activities. The latter observation is best illustrated by the frequent presence of so called "activity cliffs" (Maggiora 2006) in many chemical data sets. It is for this reason that we observed different chemical and biological neighbors for many compounds across four data sets that were evaluated. Thus, we argue that it is critical to weigh in both chemical and biological neighbor's contribution in predicting every compound's activity. Such "enhanced" aggregation underlying CBRA exploits the complementary information inherent in both the chemical and biological neighbors to arrive at the most optimal prediction.

Despite the power of relative similarity, CBRA, as with any modeling method, may still yield incorrect predictions by either or both set(s) of neighbors. In the latter case, neither biological nor chemical neighbors are instructive for model prediction or interpretation. In the former case, using either biological or chemical neighbors (instead of both as in CBRA) would yield better predictions for certain compounds. However, such accuracy provided by one set of neighbors may be limited to certain compounds and not the entire data set, as evident by the slightly smaller fraction of compounds correctly predicted by biological or chemical read-across models (50-77% overall accuracy, mean=65%, SD=9%) *vs.* that by CBRA (57-80% overall accuracy, mean=69%, SD=10%, Appendix 1 Table A1.3.1). In other words, CBRA's predictions may be less accurate than either biological or chemical read-across for a minority of compounds but the CBRA approach succeeds overall showing higher accuracy on average as compared to other read-across methods.

As with most ensemble models, the decreased interpretability and increased computational cost may outweigh the gains in accuracy (Elder 2003, Hewitt et al. 2007). CBRA, like other instance-based learners, is better suited to data sets where variable selection has already been performed to reduce noise due to irrelevant variables (Aha et al. 1991). This variable selection step is necessary because, unlike models employing variable-specific weights, all variables in CBRA, including irrelevant ones, are given equal consideration when calculating similarity.

Despite certain limitations we argue that CBRA remains transparent and interpretable since neighbors of each compound can be easily identified and important variables (chemical features or specific genes) can be elicited as illustrated in our case studies. The important variables not only suggest mechanisms of action for closer toxicological examination but may also act as markers for a particular mechanism. Such marker profiles may help to uncover the toxicity mechanisms of new compounds whose toxicities were previously unknown. Additional studies may be undertaken to investigate the potential chemical-biological relationships identified by CBRA.

In addition, the consideration of toxicokinetics may be essential for constructing a read-across argument, but CBRA does not yet take this into account in the present format. The similarity of toxicokinetic profiles, especially metabolism, is often considered before weighing similarity in terms of mechanism of action. It is therefore reasonable to conclude that the inclusion of toxicokinetic descriptors may ultimately help in improving prediction accuracy and acceptance of read-across. This is especially relevant when *in vitro* data is used to predict *in vivo* endpoints (Thomas et al. 2012). Indeed, CBRA may be extended to more than two spaces to accommodate toxicokinetic considerations although additional visualization techniques (Reif et al. 2013) may be required.

3.5.6. Recommendations for chemical-biological modeling and its application in hazard assessment

Our experience and observations with using both chemical and biological descriptors suggest the following methodological implications for predicting chemical hazards. First, biological assays such as gene expression are expected to be more predictive than more simple assays measuring binary biological responses *in vitro* (*e.g.*, binding/nonbinding to a target protein). Second, it is advantageous to consider biological variables relevant to the prediction target. The bioassays may be selected rationally according to biological pathways (Judson et al. 2011). Third, variable selection and rigorous model validation prevent selection bias towards overly optimistic models (Thomas et al. 2012). Thus, careful modeling and validation according to OECD (Q)SAR principles (OECD 2007) are necessary to ensure robust and accurate models (Tropsha 2010). Lastly, irrelevant variables may affect some classification methods more than others. For example, as explained earlier, instance-based methods including CBRA are more susceptible to irrelevant variables while others such as random forest can better tolerate noisy variables (Breiman 2001).

In addition, our work has potential practical applications for the use of read-across under REACH and other regulatory initiatives. Read-across approach guidance under REACH (http://www.reachonline.eu/REACH/EN/REACH_EN/articleXI.html) stipulates that "physicochemical properties, human health effects and environmental effects or environmental fate may be predicted from data for reference substance(s) within the group by interpolation to other substances in the group." In this sense, even though typical interpretation of "similarity" is focused on a common functional group, or common precursors/breakdown products, the biological data provides additional confirmation especially when chemical and biological data exhibit consistent patterns. CBRA's transparency in displaying the compounds selected for read-across allows users to examine the suitability of the neighbors before relying on them for subsequent prediction. As such, CBRA satisfies the requirement for "adequate and reliable documentation of the applied method" by providing a defined process for analog selection and prediction, as well as enabling visual interpretation of the similarities across several data domains.

3.6. Conclusions

Given the complex biological processes mediating chemical toxicity, hazard prediction will benefit from the inclusion of biological data in addition to chemical information. Previously, we have demonstrated that hybrid models of hepatotoxicity pooling biological (gene expression profiles) and chemical features could not achieve higher accuracy than biological models (Low et al. 2011). Herein, we have developed CBRA as an alternative method combining the same biological and chemical descriptors and demonstrated that its balanced accuracy was among the best when compared with other models using biological and/or chemical descriptors. This result was also replicated in three other data sets.

One reason for the success of CBRA is that, as a local modeling technique incorporating relative similarity weighting scheme, it relies on objective metrics to predict toxicity class of a compound when predictions made with chemical vs. biological neighbors disagree. Additionally, since prediction is based on a small number of similar compounds, both the modeling process and its outcome are transparent. Radial plots display the target compound's neighbors and their relative similarities. They allow users to examine the arguments made by the model when assigning a specific call (toxic or non-toxic) to the target compound. CBRA also highlights key biological and chemical features for further mechanistic interpretation. In summary, CBRA represents a novel hybrid read-across method that is both predictive and interpretable. It combines the simplicity and transparency of read-across methods with the benefits afforded by more sophisticated techniques such as ensemble modeling and instance-based learning while incorporating modern diverse data streams, making CBRA a potentially appealing tool for chemical hazard assessment.

CHAPTER 4. INTEGRATIVE STUDY OF ADVERSE DRUG REACTIONS: CHEMINFORMATICS PREDICTION AND PHARMACOEPIDEMIOLOGY EVALUATION OF DRUG-INDUCED STEVENS JOHNSON SYNDROME

4.1. Introduction

Adverse drug reactions (ADR) account for up to $75 billion in healthcare expenditure in the US.(Ahmad 2003, National Research Council 2007) Predicting ADR for current and investigational medications will benefit drug surveillance and minimize patient exposure to harmful drugs. Drug surveillance (or pharmacovigilance) systems currently collect and monitor spontaneous ADR reports. VigiBase, with over 12 million records, is the largest and most authoritative resource maintained by the Uppsala Monitoring Center under the auspices of the World Health Organization. Additional data concerning reported ADR can be obtained from pharmacoepidemiological (Coloma et al. 2012, Platt et al. 2012) studies, where drugs are associated with specific ADR through statistical analysis of large patient populations. Increasing digitization of health data such as electronic medical records and insurance claims provide rich sources for pharmacoepidemiology to draw from.(Hennessy 2006, Strom et al. 2012, Wilson et al. 2003)

Increasingly, cheminformatics, in particular Quantitative Structure Activity Relationships (QSAR) modeling, is used to predict ADR since they only require drug chemical structures as input (Matthews et al. 2009b, Shirakuni et al. 2012, Vilar et al. 2011, 2012). Such early ADR prediction can minimize harmful drug exposure and has established cheminformatics as an integral tool for drug development(Bender et al. 2007, Gleeson et al. 2012, Scheiber et al. 2009) and regulatory decision support (Matthews et al. 2009a,b).

Despite their common goals in predicting ADRs, pharmacoepidemiology and cheminformatics have not been used together. Herein, for the first time, we explore their combined power. Cheminformatics, using pharmacovigilance reports for building QSAR models, will uncover the relationships between drug chemical structures and ADR; thereafter, pharmacoepidemiology, using health insurance claims, will validate these models.

For this proof-of-concept study, we chose Steven Johnson Syndrome (SJS) as the ADR of interest because of its medical severity and well-established structure-activity relationships (Roujeau et al. 1995, Shirakuni et al. 2012) linking drug classes such as sulfonamide antibiotics, penicillins, and quinolones to SJS (Roujeau et al. 1995). In SJS, epithelial membranes detach, leaving denuded haemorrhagic areas with up to 30-40% mortality rate (Roujeau et al. 1995). Although the exact pathogenesis is unknown, SJS is often drug-induced and immune-mediated.(Reilly & Ju 2002) Unfortunately, many commonly used drugs have been falsely implicated with SJS (Toler & Rodriguez 2004), leading many prescribers to limit their use and hence, therapeutic options. Thus, there is a need to develop high-accuracy models capable of discriminating harmful SJS inducers and safe drugs.

Drawing from the most comprehensive data sources available (VigiBase (Lindquist 2008) ADR reports, DrugBank (Wishart et al. 2008) chemical structures, and MarketScan health insurance claims (Truven Health Analytics)), we have developed, interpreted, applied and validated QSAR prediction models of SJS (Figure 1). Models were developed using a diverse set of drugs associated with SJS according to VigiBase. Then we applied these models to virtual screening of DrugBank for potential SJS inducers, some of which were validated by pharmacoepidemiology analysis of health insurance claims. We posit that models developed in this study may not only guide rational drug design and selection of safe drug candidates but also direct pharmacovigilance surveillance of the established and emerging drugs.

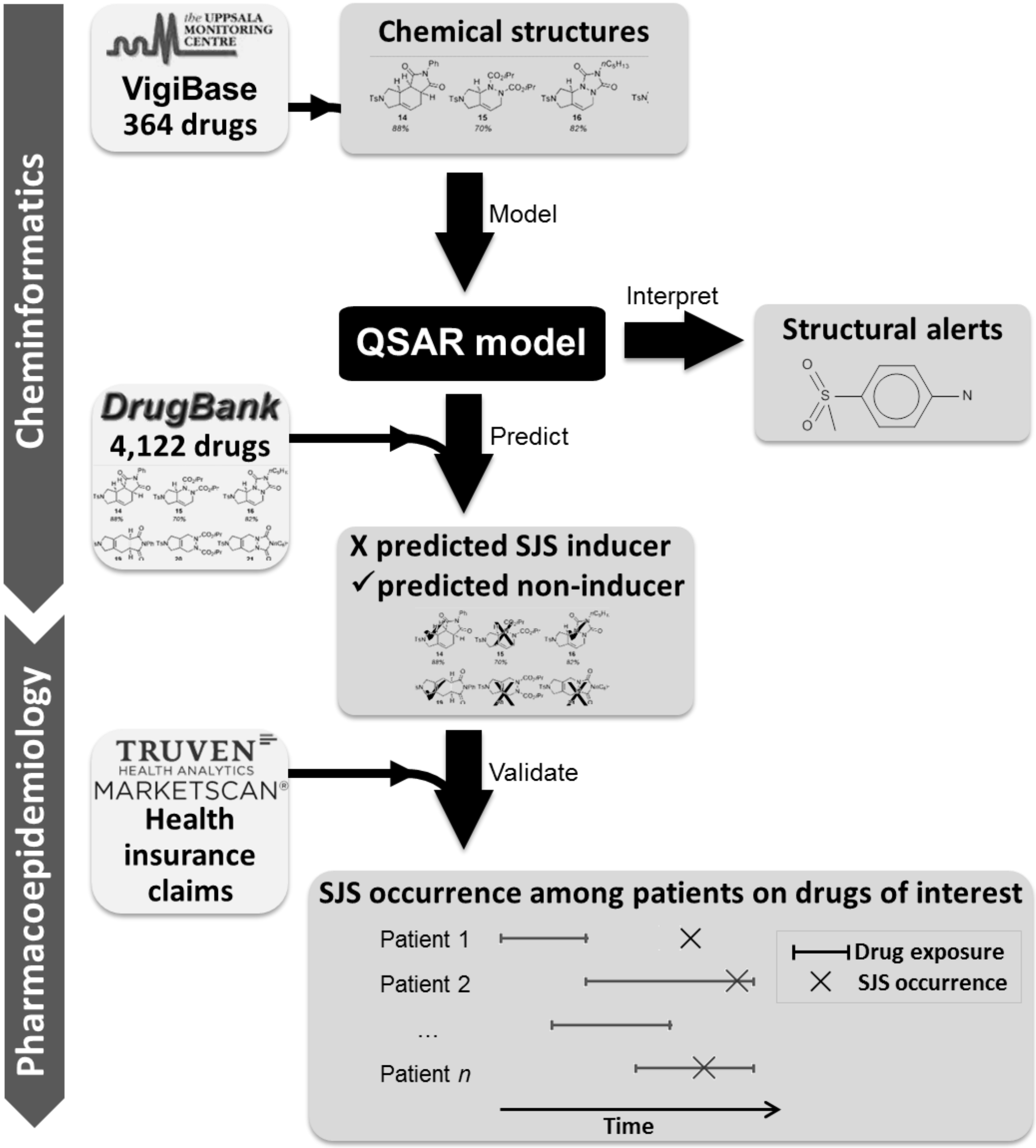


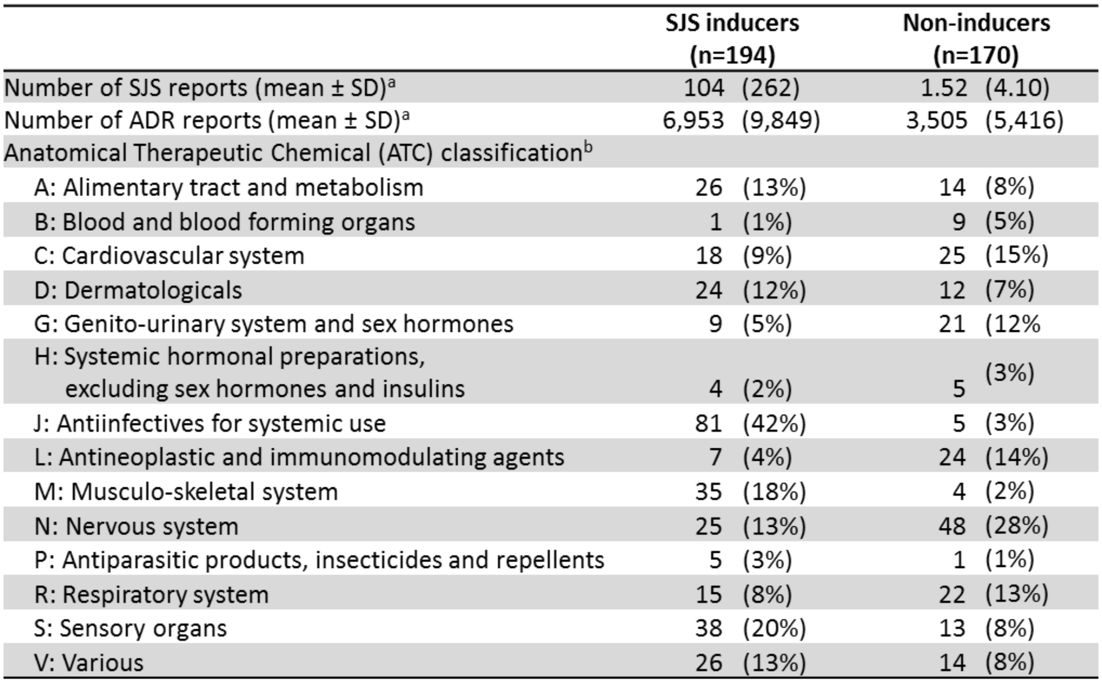
Figure 4.1. Schematic workflow bridging cheminformatics and pharmacoepidemiology. VigiBase provided 364 drugs (known SJS inducers and non-inducers) that were used for QSAR modeling. QSAR models provided structural alerts (SA) for interpretation and predicted potential SJS inducers and non-inducers in DrugBank. The predicted inducers and non-inducers were evaluated by a cohort study following patients for occurrence of SJS using health insurance claims (MarketScan).

4.2. Results

4.2.1. Properties of drugs used for QSAR modeling

A reference set of 194 SJS-inducing and 170 non-inducing drugs, defined by the disproportionate frequency of SJS spontaneous reports, were extracted from VigiBase (Table A1.4.S1). The SJS inducers (Table 4.1) had more SJS reports in VigiBase than non-inducers (mean=104 versus 1.5), more ADR reports overall (mean=6,953 versus 3,505) and were disproportionately drawn from Anatomical Therapeutic Chemical (ATC) groups J (anti-infectives) and M (musculo-skeletal system) while non-inducers were disproportionately drawn from ATC groups G (genito-urinary system and sex hormones) and N (nervous system).

Table 4.1. Properties of SJS-inducing and non-inducing drugs used for QSAR modeling



	SJS inducers (n=194)		Non-inducers (n=170)	
Number of SJS reports (mean ± SD)[a]	104	(262)	1.52	(4.10)
Number of ADR reports (mean ± SD)[a]	6,953	(9,849)	3,505	(5,416)
Anatomical Therapeutic Chemical (ATC) classification[b]				
A: Alimentary tract and metabolism	26	(13%)	14	(8%)
B: Blood and blood forming organs	1	(1%)	9	(5%)
C: Cardiovascular system	18	(9%)	25	(15%)
D: Dermatologicals	24	(12%)	12	(7%)
G: Genito-urinary system and sex hormones	9	(5%)	21	(12%
H: Systemic hormonal preparations, excluding sex hormones and insulins	4	(2%)	5	(3%)
J: Antiinfectives for systemic use	81	(42%)	5	(3%)
L: Antineoplastic and immunomodulating agents	7	(4%)	24	(14%)
M: Musculo-skeletal system	35	(18%)	4	(2%)
N: Nervous system	25	(13%)	48	(28%)
P: Antiparasitic products, insecticides and repellents	5	(3%)	1	(1%)
R: Respiratory system	15	(8%)	22	(13%)
S: Sensory organs	38	(20%)	13	(8%)
V: Various	26	(13%)	14	(8%)

[a] significant difference (p-value <0.01) by Welch t-test for unequal variances
[b] signficance was not determined for ATC as drugs could belong to multiple ATC

To explore the structure-activity landscape, a self-organizing map (SOM)(Guha et al. 2004, Kohonen 2008) clustered 364 drugs into 36 cells based on the drugs' chemical descriptors such that chemically similar drugs were placed close to one another (Figure 4.2). SOM cells were colored by the proportion of SJS inducers (pink if a majority were SJS inducers; gray if a majority were non-inducers). The emergence of an upper pink block (SJS inducer majority) and a lower gray block (non-inducer majority) suggests that the SJS inducers may be distinguished from the non-inducers based on their chemical structures, warranting further application of QSAR modeling to refine this discrimination.

Further, we examined the relationship between the drugs' chemical structures and their therapeutic uses. Each SOM cell was populated by respective ATC letters such that the letters' size and color indicated the number of drugs and proportion of SJS inducers, respectively. For instance, on the lower left corner, the largest letter "G" indicated that most of the drugs in the SOM cell belonged to ATC class G, although some belonged to other ATC classes (L, D, N). Overall, the SOM shows that non-inducers often belonged to ATC classes G and N (nervous system) as indicated by the largest letters in gray while SJS inducers often belonged to ATC classes J (anti-infectives) and M (musculo-skeletal system) as indicated by the largest ATC letters in pink. Thus, clustering with SOM allows one to draw associations among chemical structures, ATC and SJS activity.

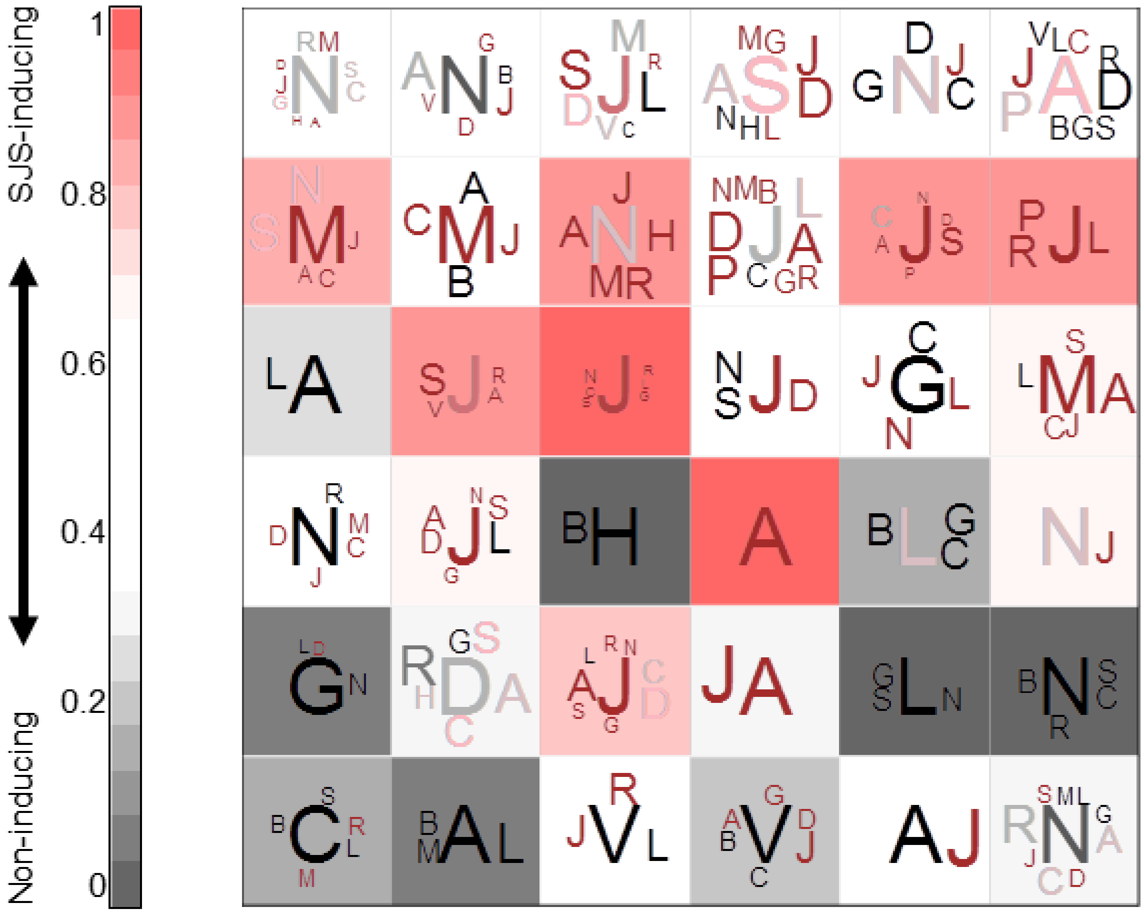


Figure 4.2. Self-organizing map (SOM) in which 364 drugs were clustered into 36 cells based on Dragon descriptor profiles. Chemically similar drugs were positioned close to one another such that their topological proximity on the SOM reflected their chemical similarity. Within each SOM cell, letters represent ATC of drugs such that the largest letters indicate the most frequent ATC. SOM cells and ATC letters were colored by their proportion of SJS inducers (pink if mostly inducers, grey if mostly non-inducers). The emergence of an upper pink block (containing mostly SJS inducers) and lower gray block (containing mostly non-inducers) suggests that inducers and non-inducers may be discriminated by chemical descriptors.

4.2.2. QSAR model performance

QSAR models were built using three sets of chemical descriptors [Dragon (Todeschini & Consonni 2000), ISIDA (Varnek et al. 2005) and Molecular ACCess System (MACCS)(Durant et al. 2002)] and two classification methods [Random Forest (RF)(Breiman 2001) and Support Vector Machines (SVM)(Vapnik 2000)]. Both the six models and their consensus (single-vote average of six predictions) showed high accuracies characterized by the Area Under the Curve (AUC) values of 75-81% (Table A1.4.S2). Coverage (i.e., the fraction of the dataset that could be reliably predicted by the models) was generally high for all models (97-100%) although a few macrolides (*e.g.,* bleomycin) were too structurally different (too large) to be predicted reliably. The *y*-randomization (i.e., random permutation of the target property) test showed that all models were unlikely to be fitted by chance (p-value < 0.05).

4.2.3. QSAR model interpretation: structural alerts (SA) for SJS

We analyzed the RF model built with the ISIDA fragments, many of which correspond to chemical functional groups. We focused on *f* most important (Strobl et al. 2008) fragments that could provide a reduced model with equivalent or lower out-of-bag (OOB)(Breiman 2001) error than that of the full model with 1,091 fragments. By examining the OOB at various *f* values, *f*=29 fragments afforded models as predictive as that built with all 1,091 fragments; thus, these 29 fragments were used for subsequent analysis to identify structural alerts associated with SJS.

Although each of these 29 discriminatory fragments could *individually* serve as an indicator for SJS activity (or lack thereof), fragments occurring frequently within the same drugs could be fused to generate larger, more accurate substructures to serve as structural alerts (SA) for SJS. To uncover these SA by co-occurrence analysis (Method 1), every pair of fragments were tested for association between their co-occurrence and SJS activity using Fisher’s exact test (Figure 4.3A). Their co-occurrences could be visualized by a network of fragment nodes, connected whenever a pair co-occurred significantly (Figure 4.3B). In the network, clusters of co-occurring fragments formed densely connected subnetworks (*i.e.*, communities) which were identified by walktrap community detection (Pons & Latapy 2005) (Figure 4.3B). Within each community, some co-occurring fragments may be assembled into a larger substructure as an indicator for either SJS class (Figure 4.3C). Of the five communities identified, two contained fragments that could be assembled into larger substructures forming SA for SJS activity (communities C1 and C2, Figures 3B-C). The first community (C1, purple) consisted of five fragments corresponding to arylamines, sulfonylarenes and sulfones that were assembled into a sulfonylarylamine SA, the substructure fitting all the fragments. The second community (blue) formed a β-lactam substructure. The green and yellow communities were composed of aliphatic chains and secondary amines that were more frequently present among non-inducers than SJS inducers, forming a “safe” substructure indicative of non-inducers (C3, C4, not shown). However, because such fragments are often present in many drugs, their practical use as “safe” substructures is limited. The remaining community (C5, pink) contained only two fragments, too small for meaningful interpretation.

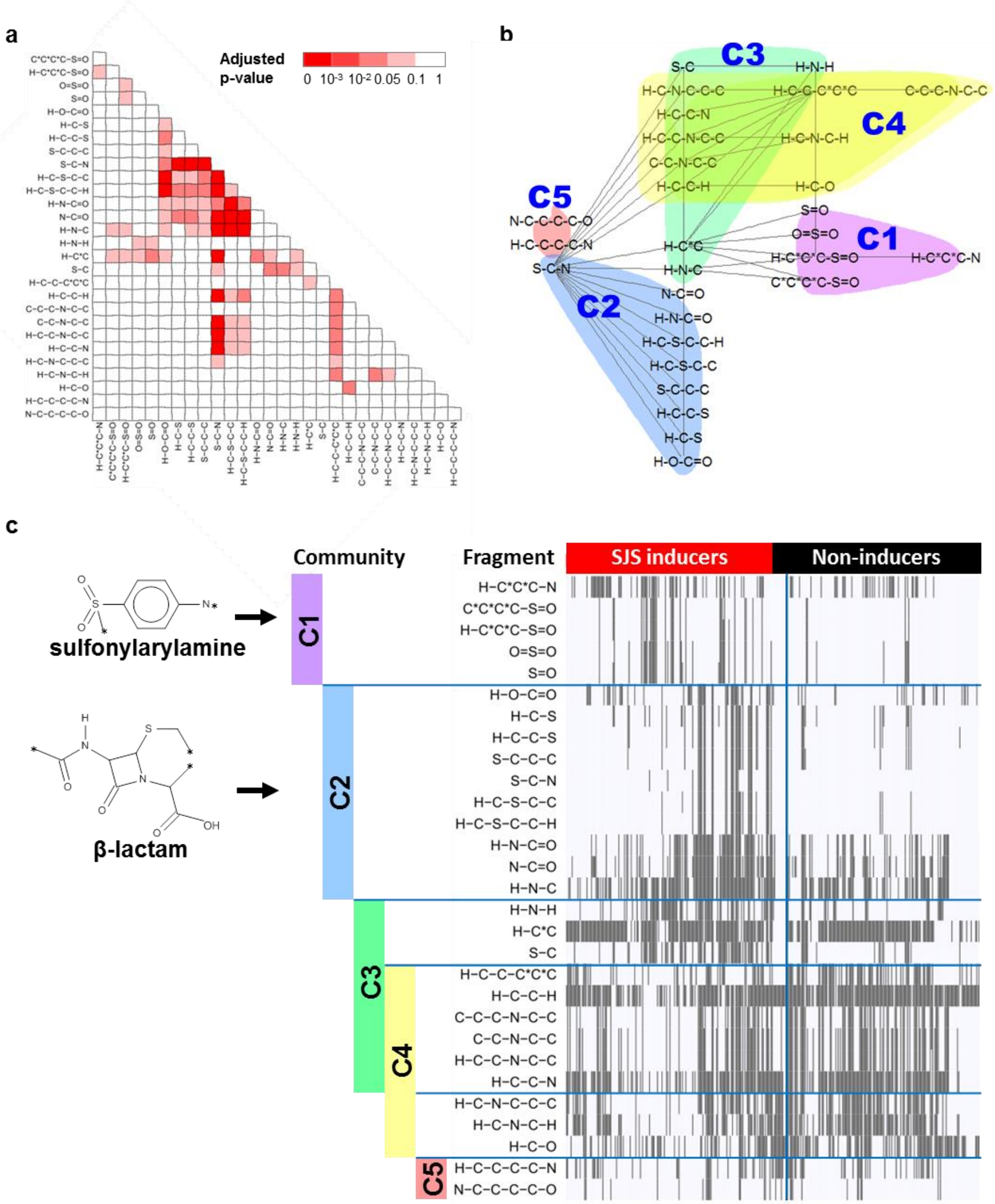


Figure 4.3. Results of co-occurrence analysis of ISIDA chemical fragments.

Figure 4.3 (continued). (a) Adjusted p-values show the association between pairwise co-occurrence of any two fragments and SJS inducing activity (from two-sided Fisher's exact test). (b) Distinctly colored communities of co-occurring fragments detected by walktrap community algorithm. Fragment nodes are connected if significantly co-occurring (p-value <0.1). (c) Heatmap shows the joint presence of co-occurring fragments within a community (e.g. purple C1, corresponding to sulfonylarylamine structural alert assembled from five co-occurring fragments, is more frequently present among SJS inducing drugs.

Both SA uncovered by the above co-occurrence analysis were consistent with those obtained from the second approach, Maximal Common Substructure (MCS), which has been commonly used to identify frequent substructures. This concordance provides evidence that co-occurrence analysis is a valid method to derive SA. However, MCS discovered two additional SA, fluoroquinolones and tetracyclines (Figure 4.4, IIIb, IVb) that were only present in at least six drugs. Because of their rarity, their key related fragments (*e.g.*, fluorinated groups, quinones) were not among the 29 most important fragments analyzed for co-occurrences. Thus, co-occurrence analysis may be better suited to detecting substructures above a minimum frequency.

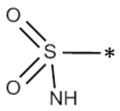
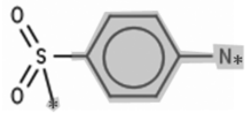
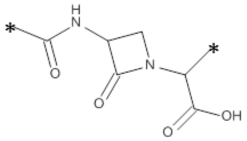
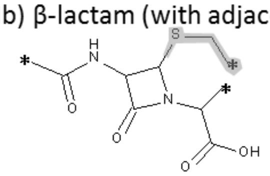
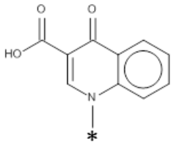
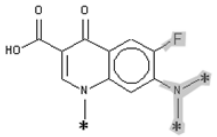
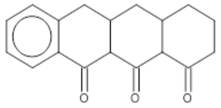
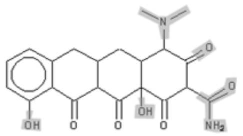
Previous structural alerts		Our structural alerts	
Ia) Sulfonamides 	29:5 (Toxic:Nontoxic) Precision=0.85	Ib) Sulfonyl arylamines 	20:0 1.00
IIa) β-lactam 	25:1 0.96	IIb) β-lactam (with adjacent sulfur) 	24:0 1.00
IIIa) Quinolone 	6:1 0.86	IIIb) Fluoroquinolone 	6:1 0.86
IVa) tetracyclines/ anthracyclines 	6:3 0.67	IVa) tetracycline 	6:0 1.00

Figure 4.4. Structural alerts (SA) whose presence in a drug alerts for SJS inducing activity. Left column shows previously inferred substructures. Right column shows SA uncovered in this study. Structural differences are highlighted in gray.

Substructures previously associated with SJS inducers (Roujeau et al. 1995, 2011) are shown in Figure 4.4. Such substructures were inferred from drug classes implicated with SJS such as sulfonamide antibiotics, penicillins, quinolones and tetracyclines (Roujeau et al. 1995, 2011). Our systematic chemical analysis found larger, more specific substructures (Figure 4.4, right column) that were more likely to yield true positives (i.e., higher precision). For example, the sulfonylarylamine SA (Figure 4.4, Ib) correctly identified drugs causing SJS all 20 times it was present in a drug, unlike the sulfonamide SA (Figure 4.4, Ia) which falsely predicted the SJS-inducing potential for some sulfonamide drugs. In minimizing false positives, more precise SA could “spare” drugs from wrongful association with SJS and leave more drug options available for use.

4.2.4. QSAR model application: predict SJS inducers and non-inducers in DrugBank

We used the best QSAR model (Dragon-RF) for virtual screening of DrugBank, assessing 4,122 drug structures for potential SJS activity (Appendix 1 Table A1.4.S4). Among the ten most likely SJS inducers (excluding experimental drugs), eight contained either the sulfonylarylamine or β-lactam with adjacent sulfur SA (Appendix 2 Figure A2.4.S2). Among the ten most likely non-inducers, etonogestrel, mestranol and rapacuronium were chemically similar to many steroidal non-inducers in our reference set such as progesterone.

4.2.4.1. Validation of predictions by QSAR models

4.2.4.1.1. By VigiBase, ChemoText and Micromedex 2.0.

We checked VigiBase and the medical literature (ChemoText(Baker & Hemminger 2010) and MicroMedex 2.0) for reports of SJS associated with the predicted SJS inducers and non-inducers (Table 4.2). Between predicted inducers and predicted non-inducers, the former were associated with disproportionally more SJS reports and higher Information Component (IC)(Bate et al. 1998) values indicative of higher-than-expected SJS reporting in VigiBase, and more instances of SJS in ChemoText and Micromedex 2.0. Despite the evidence in VigiBase, these predicted drugs were not included in our reference set for modeling as they were not obvious candidates for known inducers and non-inducers due to co-reporting with other co-medications and low usage (evident by the few ADR reports). Where ADR reports were few, we cautioned against relying on IC as the only evidence of

SJS given that their 95% credibility intervals were very wide. Nevertheless, the general trends in the IC values and other data sources showed that the predicted inducers were associated with more SJS instances than predicted non-inducers in support of our models' predictions.

Table 4.2. Most likely SJS inducers and non-inducers in the DrugBank, as predicted by Dragon-RF model

				VigiBase			ChemoText	Micromedex
DrugBank ID	**Predicted Value**	**SD**	**Name**	**SJS reports**	**All ADR reports**	**IC[a]**	**SJS articles[b]**	
Predicted inducers (from DrugBank)								
DB01581	0.978	0.010	Sulfamerazine	0	1	-0.01	2	N
DB01332	0.967	0.006	Ceftizoxime	2	748	-0.26	0	Y
DB00493	0.966	0.016	Cefotaxime	40	7550	0.66	15	Y
DB00576	0.964	0.007	Sulfamethizole	5	490	1.37	5	Y
DB01325	0.963	0.007	Quinethazone	0	25	-0.22	0	N
DB00880	0.959	0.040	Chlorothiazide	2	800	-0.34	1	Y
DB00891	0.955	0.011	Sulfapyridine	0	29	-0.25	2	N
DB01333	0.951	0.012	Cefradine	3	994	-0.12	1	N[c]
DB00301	0.937	0.020	Flucloxacillin	29	5272	0.71	3	N
DB01607	0.937	0.006	Ticarcillin	2	338	-0.11	0	Y
Predicted non-inducers (from DrugBank)								
DB00294	0.066	0.031	Etonogestrel	1	4443	-3.35	0	N[c]
DB01357	0.084	0.036	Mestranol	0	26	-0.23	2	N
DB00202	0.099	0.144	Succinylcholine	4	3581	-1.46	0	N
DB01160	0.100	0.080	Dinoprost	0	161	-1.05	0	N
DB01088	0.103	0.020	Iloprost	1	1518	-1.88	0	N
DB01049	0.109	0.037	Ergoloid mesylate	2	171	1.23	0	N
DB00966	0.110	0.023	Telmisartan	8	4845	-0.96	0	N[c]
DB01089	0.120	0.044	Deserpidine	0	9	-0.08	0	N
DB04834	0.123	0.079	Rapacuronium	0	112	-0.80	0	N
DB00654	0.124	0.042	Latanoprost	3	5423	-2.40	1	N[c]
Known inducers/positive controls (from reference set)								
-	-	-	Sulfamethoxazole	37	971	1.43	104	Y
-	-	-	Amoxicillin	648	48501	1.36	44	Y
Known non-inducers/negative controls (from reference set)								
-	-	-	Progesterone	0	3825	-4.72	0	N[c]
-	-	-	Vardenafil	0	4506	-4.95	0	N

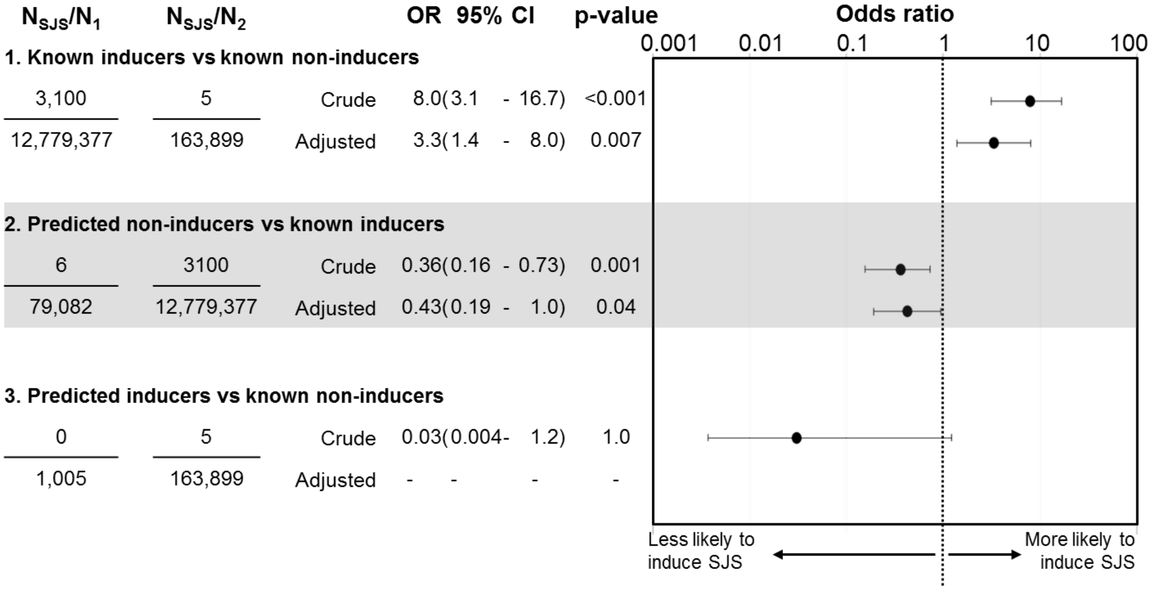
[a] Information component (IC) is a disproportionality frequency measuring the number of SJS reports lower than or higher than expected in VigiBase.

[b] Number of articles in Medline matching the search terms: [drugname] AND "Stevens-Johnson Syndrome"[Mesh] OR "Erythema Multiforme"[Mesh] OR "epidermal necrolysis, toxic"[MeSH] Filters: Humans (as of Februrary 2013)

[c] Hypersensitivity reaction although SJS was not explicitly mentioned

4.2.4.1.2. By pharmacoepidemiology analysis of health insurance claims.

Pharmacoepidemiology evaluation of MarketScan (Truven Health Analytics) health insurance claims data found predicted non-inducers as a group were associated with lower odds of SJS compared to known inducers (adjusted odds ratio, OR=0.43, 95% CI [0.19, 1.0], p-value=0.04, Figure 4.5). Comparisons involving predicted inducers were underpowered and could not be evaluated as there were only 1,005 patients on predicted inducers, too few for even one observable SJS case.



N_{SJS}/N_1	N_{SJS}/N_2		OR 95% CI	p-value
1. Known inducers vs known non-inducers				
3,100	5	Crude	8.0(3.1 - 16.7)	<0.001
12,779,377	163,899	Adjusted	3.3(1.4 - 8.0)	0.007
2. Predicted non-inducers vs known inducers				
6	3100	Crude	0.36(0.16 - 0.73)	0.001
79,082	12,779,377	Adjusted	0.43(0.19 - 1.0)	0.04
3. Predicted inducers vs known non-inducers				
0	5	Crude	0.03(0.004- 1.2)	1.0
1,005	163,899	Adjusted	- - -	-

Figure 4.5. Crude and adjusted odds ratio (with 95% confidence intervals) comparing patients on SJS inducers (known or predicted) vs those on non-inducers (known or predicted).

4.3. Discussion

To meet our objectives of developing, interpreting, applying and validating QSAR models of SJS, we first developed QSAR classifiers that predicted SJS inducers and non-inducers from chemical structures with consistently high accuracy (75-81% AUC, Appendix 1 Table A1.4.S3).

Second, we interpreted the ISIDA models by identifying the most predictive fragments from which we further identified co-occurring combinations (SA) frequently associated with the inducer class (Figure 4.3). Although the SA were present in fewer drugs, they alerted for SJS activity with greater precision than previously suspected substructures, defining an enriched chemical space for which their presence more effectively alerted for SJS (Figure 4.4).

The additional chemical features encapsulated in our larger SA offered mechanistic clues. For example, sulfonamides antibiotics have long been implicated with SJS (Roujeau et al. 1995) although it is known that sulfonamides alone do not induce SJS (Toler & Rodriguez 2004). Instead, studies have attributed immunogenic reactions related to SJS to an arylamine group within the sulfonylarylamine (Brackett et al. 2004) SA (Figure 4.4, Ib). The purported mechanism involves the metabolic transformation of the arylamine group into a reactive nitroso metabolite which covalently binds to cellular macromolecules to initiate an immune response consistent with the hapten hypothesis (Brackett et al. 2004, Naisbitt et al. 1999, Toler & Rodriguez 2004). Arylamines are generally rare among drugs due to their reactivity. Exceptions are drugs such as sulfonamide antibiotics, which contain a sulfone group (SO_2) in the electron-withdrawing *para*-position to stabilize the arylamine against overt toxicity but not exculpate it from metabolizing into the nitroso culprit (Uetrecht 2002).

The other SA, β-lactam with adjacent sulfur (Figure 4.4, IIb), suggests that the additional sulfur atom may be necessary for SJS activity. By specifying the adjacent sulfur atom, precision increased to 100% such that all β-lactam antibiotics containing it were SJS inducers. Conversely, analogs without the adjacent sulfur atom such as latamoxef were non-inducers.

Our third SA refers to a fluoroquinolone (Figure 4.4, IIIb) instead of quinolone as previously suspected. However, because all the quinolones in our study were also fluoroquinolones, we could not compare them and conclude that one was a better alert. We note that such a distinction between the two may be irrelevant as most unfluorinated quinolones have been discontinued in favor of the more efficacious fluoroquinolones(King et al. 2000).

Our fourth SA refers to tetracycline antibiotics instead of the more general four-ring system present in both tetracycline antibiotics and anthracyclines. In our study, all six tetracycline antibiotics were inducers while all three anthracyclines were non-inducers. Their structural differences lie in the presence of a dimethylamine group and absence of a sugar ring in tetracycline antibiotics. In using a more refined SA that can differentiate the SJS-inducing tetracycline antibiotics from the non-inducing anthracyclines, we improved the precision to 100%.

Other substructures such as the aromatic ring has been suggested as a SA for anticonvulsants by a previous study (Handoko et al. 2008). However, we did not find this trend in our study using our expanded set of drugs including non-anticonvulsants. One reason may be the ubiquity of the aromatic ring in both SJS inducers and non-inducers.

Third, we demonstrated the practical utility of the QSAR models, using them to screen the DrugBank library of 4,122 drug structures for potential SJS inducers. Those predicted with high confidence were chemically similar to drugs in our reference set used for training QSAR models (Appendix 2 Figure A2.4.S2).

Fourth, when verified against SJS reports, predicted inducers were associated with more SJS reports than predicted non-inducers (Table 4.2). Pharmacoepidemiology evaluation also found predicted non-inducers associated with lower SJS odds (Figure 4.6). However, similar assessment of predicted inducers was inconclusive as both predicted inducer usage and the SJS outcome were rare.

Some pitfalls of our approach warrant a discussion. One limitation stems from VigiBase's voluntary spontaneous reporting system which is prone to underreporting and reporting bias. Nevertheless, it remains the largest source of ADR reports providing the largest set of reference drugs for analysis. Another vulnerability relates to the definition of SJS inducers and non-inducers for modeling. Drugs were statistically defined by disproportionality analysis (see Methods) instead of being clinically defined by a gold standard. Fortunately, the two definitions have been reported to be in good agreement (83% accuracy) (Harpaz et al. 2013). Another drawback is the limited predictivity by chemical structures alone (up to 81% AUC) despite rigorous modeling and validation in line with OECD QSAR guidelines (OECD 2007). Better understanding of the SJS mechanisms may help to identify non-chemical factors such as metabolism (Hou & Wang 2008) currently unexplained by the chemical descriptors used. The pharmacoepidemiology analysis of insurance claims posed statistical limitations for rare outcomes such as SJS. For lack of temporal data, we could not demonstrate earlier detection of ADR than current methods. This could be addressed by a prospective validation in which later data would be set aside for validating models built on earlier data.

4.4. Conclusions

In conclusion, our tiered strategy demonstrated the combined power of cheminformatics and pharmacoepidemiology in predicting and validating drugs that could cause SJS. We advocate applying such a strategy for other ADRs.

4.5. Methods

4.5.1. VigiBase as data source.

For QSAR modeling, a reference set of drugs was extracted based on their reporting correlations with SJS in VigiBase (~20,000 drugs and 2,000 ADR among 7,014,658 reports from 107 countries as of February 2012, coded according to WHO Drug Dictionary Enhanced and WHO-ART(Lindquist 2008)).

SJS inducers were drugs with higher-than-expected reporting with SJS (i.e., positive coefficients in a shrinkage regression model (Caster et al. 2010)). Non-inducers were defined by the following criteria. For drugs with <1000 ADR reports in total, the criterion was to never be reported with SJS. For drugs with ≥1000 reports in total, both these criteria were required: (i) no reports where the drug was the only drug suspected of causing SJS; and (ii) disproportionately few SJS reports (i.e., negative IC 95% credibility interval (Bate et al. 1998, Norén et al. 2011)).

4.5.2. Chemical structures.

Chemical structures were retrieved for drugs excluding mixtures and biologics. Chemical curation ensured that drug structures were correctly represented and standardized prior to the modeling(Fourches et al. 2010). After removing salts, metal-containing compounds, large molecules (molecular weight > 2,000) and structural duplicates (ChemAxon v.5.0; Pipeline Pilot Student Edition v.6.1.5, Accelrys Inc), 194 SJS inducers and 170 non-inducers remained for QSAR modeling (Appendix 1 Table A1.4.S1).

Molecular structures were converted into three different types of chemical descriptors: 457 Dragon descriptors, 3,404 ISIDA fragments, and 166 MACCS fingerprints.

The 457 Dragon descriptors (v.5.5, Talete SRL, Milan, Italy)(Todeschini & Consonni 2000) included constitutional groups, functional groups, atom-centered fragments, molecular properties and 2D frequency fingerprints. ISIDA/Fragmentor (Varnek et al. 2005) split each chemical structure into substructural fragments containing 2 to 6 atoms in linear sequence. Fragment descriptors were binarized depending on whether they were present (1) or absence (0) in the drug. MACCS fingerprints are binary representations of a predefined set of 166 chemical features (Durant et al. 2002).

All subsequent analyses were performed in R (v.2.14). Continuous descriptors (Dragon) were autoscaled to z-scores. Descriptors were excluded if they were invariant (< 0.001 standardized standard deviation, > 99% constant values) or intercorrelated (if pairwise $r^2 > 0.99$, randomly remove one of the two descriptors), such that 354 Dragon, 138 MACCS and 1,091 ISIDA descriptors remained for modeling (Appendix 1 Table A1.4.S2).

4.5.3. Preliminary chemical exploration by SOM clustering of drugs in chemical space

To visualize the 364 drugs in the chemical space, they were clustered by their Dragon descriptor profiles using a SOM(Kohonen 2008) which projected the drugs onto a two-dimensional 6 x 6 grid of cells such that similar drugs were clustered together within a cell and cells with similar groups of drugs were placed closed to one another (Figure 4.2). Thus, the topological distance between the drugs based on SOM reflects the Euclidean distance of their chemical descriptor profiles. Each cell was colored by its proportion of SJS inducers (gray if mostly SJS inducers, pink if mostly SJS inducers). Each cell was also populated with letters representing the drugs’ ATC codes, and the letters were sized and colored according to the number of drugs and proportion of SJS inducers.

4.5.4. QSAR model development.

For each of the three descriptor sets, two classification methods, RF (Breiman 2001) and SVM (Vapnik 2000), were used to build QSAR models. All models were evaluated by external 5-fold cross validation (CV) (Tropsha & Golbraikh 2007) whereby the data set was divided randomly into five equal parts and each individual part was systematically used as an external validation set while the remaining 80% was used as a modeling set. Additionally, modeling parameters were tuned by internal 5-fold CV which further split each modeling set. A tuned model had modeling parameters that resulted in the smallest error averaged across the five training sets from internal 5-fold CV. This tuned model was subsequently externally validated with the corresponding external set, which was never used for parameter tuning. Only prediction accuracies for the external sets were reported (Appendix 1 Table A1.4.S3).

Models were assessed by specificity, sensitivity, balanced accuracy, AUC and coverage. Balanced accuracy is the average of specificity and sensitivity. Coverage is the fraction of drugs in the external set that are within the models' extrapolation domain (Tropsha et al. 2003). Additionally, precision was used to assess the SA. Standard errors of all metrics were calculated by bootstrapping (Efron & Tibshirani 1986) with 1000 trials.

We performed *y*-randomization to ensure that models were robust and not due to chance correlations.(Wold & Eriksson 1995) After permuting the *y* activity labels in the modeling sets, models were rebuilt following the same procedures as outlined above for non-permuted data. This process was repeated 30 times to generate a distribution of *y*-randomized model accuracies for comparison under a one-tailed one-sample t-test.

4.5.5. QSAR model interpretation.

Model interpretation involved identifying key chemical predictors of SJS in terms of *f* most important individual chemical fragments and fused substructures reconstituted from the fragments. This was only possible for the models based on ISIDA fragments.

4.5.5.1. *f* most important chemical fragments.

From the ISIDA-RF model, ISIDA fragments were ranked by RF conditional importance (Strobl et al. 2008). Because the fragment ranking varied slightly across the five models generated with 5-fold external CV, only *f* fragments that were consistently among the top 10, 25, 50, 75, and 150 fragments in all five models were selected (see Appendix 3). We then determined the smallest number of *f* fragments that could yield RF models with out-of-bag error that was smaller or equal to that of full RF models with 1,091 fragments (Appendix 2 Figure A2.4.S1).

4.5.5.2. Structural alerts by Method 1 (co-occurring fragments)

The fused SA were reconstituted from clusters of fragments that co-occurred more frequently in SJS inducers than in non-inducers. All possible pairs of *f* fragments were tested for higher-than-expected co-occurrence in inducers than in non-inducers by a two-tailed Fisher's exact test. A fragment pair was said to have significant co-occurrence when its p-value was <0.1 (after adjustment for multiple testing by permutation; see Appendix 3 and Figure 4.3A).

Co-occurring fragments were represented by a network where fragment nodes were connected if they co-occurred significantly (Figure 4.3B). Frequently co-occurring fragments formed densely connected subnetworks known as communities in network analysis. Communities were detected by the walktrap algorithm, a bottom-up approach which stochastically agglomerated the fragment nodes such that they were disproportionately more connected with nodes inside the community than outside.(Pons & Latapy 2005) Within a distinctly colored community, the fragments are said to co-occur more frequently in drugs of one class *vs*. the other. Hence, they can be assembled into a larger substructure as SA for a certain class of drugs.

Structural alerts by Method 2 (maximal common substructures, MCS).

MCS (Chakravarti et al. 2012) provided a second set of SA for comparison with those obtained by co-occurring fragments (Method 1). MCS extracted the largest substructures more frequently associated with SJS inducers than with non-inducers that would meet all the following pre-defined criteria: size ≥8 atoms, frequency ratio ≥2 and derived from ≥6 molecules.

4.5.6. QSAR model application: predict SJS inducers and non-inducers in DrugBank.

As an application of our best QSAR model (RF model of Dragon descriptors), we screened the DrugBank library of 4,122 drugs (Wishart et al. 2008) for potential SJS inducers, after excluding drugs used for modeling and applying the same chemical curation and descriptor treatment procedures used earlier for modeling.

4.5.7. Validate predictions by QSAR models.

The following drugs were evaluated for evidence of SJS (or lack of) using VigiBase (Lindquist 2008), ChemoText (Baker & Hemminger 2010), Micromedex 2.0 and MarketScan (Truven Health Analytics) (Table 4.2): top ten predicted SJS inducers, top ten predicted non-inducers, two known inducers (*i.e.*, positive controls) and two known non-inducers (*i.e.*, negative controls). Drugs were ranked according to the consistency of predictions by the five models stemming from 5-fold CV.

4.5.7.1. Validation by VigiBase, ChemoText and Micromedex 2.0.

For each drug of interest, we queried VigiBase for the IC (Bate et al. 1998) disproportionality measure of SJS reports. From ChemoText, a chemocentric database of MeSH annotations sourced from PubMed (Baker & Hemminger 2010), we counted the number of human studies co-annotating the drug of interest and “Stevens-Johnson syndrome”, “erythema multiforme”, or “epidermal necrolysis, toxic”, inclusive of their MeSH synonyms. From Micromedex 2.0, we looked for mention of SJS and related hypersensitivity.

4.5.7.2. Validation by pharmacoepidemiology analysis of health insurance claims.

We performed a retrospective cohort study that followed patients prescribed drugs predicted as inducers or non-inducers for occurrences of SJS using MarketScan (Truven Health Analytics), the largest database of US health insurance claims. Eligible patients on the drugs of interest (in either outpatient and inpatient setting) between 2000 and 2011 were selected according to an incident user study design (Ray 2003) using the following criteria: all ages, washout period ≥30 days before index drug date and follow-up period of 45 days after index date (see Appendix 3 and Appendix 2 Figure A2.4.S3). The short washout period (≥30 days) boosted the number of eligible patients on several rare drugs. The short follow-up period (45 days) was chosen to reflect the quick onset of drug-induced SJS, typically expected within 21 days of initial drug use (Chan et al. 1990). During the study period, patients must have maintained continuous insurance coverage (<7 days uninsured). Drug exposure period started from index drug date to the end of the days’ supply of the last prescription fill, allowing a gap of ≤30 days between fills. Patient characteristics considered included age, sex, US geographical region, employment status and employment industry (Appendix 1 Table A1.4.S5). Cases of SJS were defined as patients with any outpatient or inpatient diagnosis coded 695.1x according to the International Classification of Diseases, Ninth Revision (ICD-9) as used by previous studies (Chan et al. 1990, Eisenberg et al. 2012, Hällgren et al. 2003, Schneider et al. 2012, Strom et al. 1991a,b).

For the comparisons between inducers and non-inducers, we computed crude OR and OR obtained from logistic regression adjusted for age and sex. Crude OR was calculated with small-sample adjustment (Jewell 1986) to account for the small cell counts (≤5).

CHAPTER 5. GENERAL DISCUSSION AND CONCLUSIONS

This dissertation presented several integrative approaches to address the problems facing toxicity prediction models, namely the lack of accuracy and interpretation due to incomplete formulation of chemical and biological factors central to chemical toxicity. To this end, integrative chemical-biological modeling combining both chemical and biological factors enhanced interpretation but not predictivity (Chapter 2). In efforts to further improve predictivity, chemical-biological read-across (CBRA) was developed (Chapter 3). Besides integrating chemical structures and bioassays, the use of alternative data sources such as patient health insurance claims may increase the relevance of our models to human toxicity. Chapter 4 demonstrates a feasible workflow coupling cheminformatics with pharmacoepidemiology to predict and validate drugs likely to induce Stevens Johnson Syndrome (SJS).

Note that since the conclusions, study limitations, and future perspectives specific to each chapter have been described in detail in the chapters themselves, this chapter will instead summarize the key findings, discuss their contributions, highlight study limitations and propose recommendations for future work.

5.1. Summary of key findings

5.1.1. Integrative chemical-biological modeling with existing methods improved interpretability but not predictivity (Chapter 2)

Pooling chemical structures and toxicogenomics assays for modeling with existing machine learning methods (*k*NN, SVM, RF, DWD) did not improve prediction performance as expected. However, chemical and toxicogenomic markers predictive of hepatotoxicity were identified. Instead of models based on over 30,000 genes, 85 gene markers could predict hepatotoxicity with 76% balanced accuracy (Section 2.5). When mapped onto pathways, the 85 genes signaled changes in ER stress and mitochondrial regulation, underscoring their importance in mitigating hepatotoxicity. Structural alerts suggested metabolic activation as a major mechanism underlying hepatotoxicity. The liver, being a primary site for metabolism, converts many drugs into reactive metabolites capable of eliciting oxidative stress.

The lower-than-expected prediction performance was a major shortcoming which prompted the subsequent development of an integrative method that would optimize the use of both chemical structures and bioassays for toxicity prediction (Chapter 3).

5.1.2. Novel integrative chemical-biological read-across (CBRA) improved predictivity and interpretability (Chapter 3)

Using CBRA to integrate chemical structural and bioassay assay in four data sets resulted in models with consistently high prediction performance compared to other models using either or both data types. CBRA learns from both chemical analogs and biological analogs for toxicity prediction and does so in a way that exploits the complementary information between them by an appropriate similarity-weighted aggregate.

CBRA is conceptually simple and lends itself to visualization and automation, making it an appealing tool for high-throughput regulatory assessment. To attain transparency, CBRA is represented as a radial plot which visually displays the most similar chemical and biological analogs such that the user can assess their suitability for read-across. To attain interpretability, CBRA calculates a feature importance score for identifying important chemical and biological features such that chemical and biological insights may be drawn.

CBRA also highlights important chemico-biological associations that would have been missed in chemical-only or biological-only analysis. In the analysis of benzbromarone (Section 3.5.4), CBRA uncovered a link between bent molecular structures and PPAR activity. Chemical analogs (e.g. ethinyl estradiol) resembling the bent molecular structure of benzbromarone were hypothesized to share benzbromarone's PPAR activity. Such concurrent study of chemical and biological features generates testable hypotheses for further research inquiry.

5.1.3. Cheminformatics prediction and pharmacoepidemiology validation demonstrated a practical tiered approach for detecting adverse drug reactions (Chapter 4)

Using the most comprehensive data sources available, QSAR models were developed to predict drug-induced SJS. Using the interpretation framework developed in Chapter 4, the QSAR models were deciphered for key chemical features associated with SJS. To demonstrate the practical utility of the QSAR models, the DrugBank library of 4,122 drugs was screened for SJS. Predicted SJS inducers and non-inducers were subsequently validated by a pharmacoepidemiology analysis of health insurance claims which confirmed predicted non-inducers to be associated with lower odds of SJS.

Out of 1,091 chemical fragments, the QSAR model identified 29 chemical fragments necessary for predicting drug-induced SJS. Structural alerts, generated by a co-occurrence analysis of the 29 fragments, predicted SJS with greater precision than previous structural alerts (Section 4.5). The structural alerts also hinted at key functional groups possibly leading to SJS. For instance, our sulfonyl arylamine structural alert found the arylamine group to be a critical feature for SJS. The arylamine group, being relatively reactive, is known to form immunogenic protein adducts inciting a SJS response (Naisbitt et al. 1999).

Notably, this study demonstrated a feasible scheme for coupling cheminformatics and pharmacoepidemiology for pharmacovigilance detection. QSAR models predicted high risk drugs from DrugBank while subsequent in-depth pharmacoepidemiology validated the results. Such a tiered approach combines the high throughput advantage of cheminformatics with the statistical rigor of pharmacoepidemiology while avoiding the ethical and time obstacles of conducting additional clinical trials.

5.2. Contributions and practical implications

5.2.1. Improved predictivity

Current fragmented efforts in predicting toxicity from only biological or chemical considerations have not maximized the opportunities enabled by the new toxicity data landscape of deeper biological assay characterization and broader chemical scope. More importantly, the models do not fully reflect the underlying toxicological processes which result from a complex interplay between the chemical inducer and biological host.

Chapters 2 and 3 exemplified attempts to reconcile the chemical and biological domains for predicting hepatotoxicity. When data pooling of chemical structures and biological assays failed to increase prediction performance of models as expected, an integrative method, CBRA, was developed to exploit the complementary information between chemical structures and biological assays. It acknowledges that chemical structures and biological assays are not always equally predictive of toxicity in every compound and thus, allows for different weights depending on the chemical and biological neighbors used for learning. In using a similarity weight, learning from similar analogs is favored over that from dissimilar analogs. As a result, previous conflicts between chemical-based and biological-based predictions were resolved by CBRA, netting an overall gain in the number of compounds correctly predicted. This was shown in four data sets where CBRA was consistently among the best models (Figure 3.6).

5.2.2. Improved interpretability

Interpretable models that can highlight key chemical and biological features among thousands allows us to focus our testing resources, deepen our understanding of the toxicological processes and open up new lines of inquiry for further experimentation. To this end, Chapter 2 identified 85 gene markers which were predictive of hepatotoxicity and were indicative of mechanisms related to aberrant liver growth and repair.

A novel method of deriving structural alerts based on co-occurring chemical fragments was presented in Chapter 4. Unlike current methods based on expert opinion or substructure mining, this new method drew upon the chemical descriptors used in the QSAR model, increasing the model’s explanatory power and validity.

Structural alerts, as heuristic indicators, allow a user to quickly estimate toxicity by their presence in a compound without the need for thorough modeling. Particularly, they often correspond to chemical functional groups and are highly intuitive to a chemist who can then avoid such substructures during molecular design. The increased precision of our structural alerts for SJS meant that fewer drugs were falsely implicated with SJS (i.e. fewer false positives), making more drugs available for use. In contrast, previous cautious avoidance of all sulfonamides has left prescribers and patients with fewer drug options (Toler & Rodriguez 2004). Besides increased precision, the structural alerts also suggested chemical mediators of drug-induced SJS such as the sulfoynyl arylamine group whose role in SJS has been reported in (Naisbitt et al. 1999).

CBRA, the integrative chemical-biological method developed in Chapter 3, achieved predictivity without compromising interpretability. Transparent and highly visual, CBRA is in line with current risk assessment efforts such as ToxPi (Reif et al. 2013), enabling large-scale automation while illuminating the assessment process. Not only does CBRA display the most similar chemical and biological analogs for read-across, it identifies key chemical and biological features and potential associations between them for subsequent inquiry.

5.2.3. Practical integrative schemes for toxicity prediction

This dissertation demonstrated the feasibility of integrating data and techniques from cheminformatics, bioinformatics and epidemiology for the study of chemical toxicity. Presented here are two ways: 1) using cheminformatics and bioinformatics (specifically toxicogenomics) for toxicity prediction, and 2) coupling cheminformatics prediction with pharmacoepidemiology validation. The first way recognizes that chemical toxicity arises from the complex interactions between the chemical toxicant and the biological host and tries to account for both chemical and biological factors during toxicity prediction. To this end, several integrative chemical-biological approaches were explored and compared in Chapter 3. We discovered that the naïve use of existing classification methods with data pooling or model pooling did not always result in improved predictivity, requiring new methods such as CBRA to be developed.

The second way combined the advantages of two disciplines, the high throughput scale of cheminformatics models and the statistical rigor of pharmacoepidemiology, to address some of the inadequacies faced by current pharmacovigilance efforts. Spontaneous reports, the primary source of data for ADR detection, are known to underreport and are prone to reporting bias. Increasingly, complementary data sources such as chemical structures, biological assays and clinical records are drawn upon in order to detect ADR more quickly and accurately. The study of clinical records has put the spotlight on pharmacoepidemiology which provides the tools to reliably link human health effects to drug exposure (or chemical exposure in the closely related branch of environmental epidemiology). As human experiments are not always feasible or ethical, the use of (pharmaco)epidemiology for assessing human toxicity becomes ever more important, especially when animal or *in vitro* models extrapolate poorly to human toxicity. However, because (pharmaco)epidemiology requires careful study design and rigorous statistical adjustment of confounders, it is less amenable to large-scale automated analysis unlike cheminformatics.

To overcome the lower throughput scale of pharmacoepidemiology, Chapter 4 presented a practical tiered approach in which high throughput cheminformatics first predicted from chemical structures drugs likely to induce SJS. Next, likely inducers were validated by pharmacoepidemiology analysis of health insurance claims of over 13 million patients. Such a tiered approach meets the goals of the new proposal for active drug surveillance (e.g. Mini-Sentinel) to apply increasingly rigorous methods to elicit true signals of ADR (Platt et al. 2012).

5.3. Limitations and future solutions

5.3.1. Chemical data limitations and future solutions

QSAR modeling is inherently limited by the chemical space representation of the chemicals available for modeling. Chemical space is defined as the multi-dimensional descriptor space spanned by the chemicals within a data set. An ideal chemical space for QSAR modeling should be densely and uniformly populated with sufficient compounds and be wide enough to be applicable to most test compounds. Such qualities allude to the chemical diversity and representativity of the data set (Bayada et al. 1999). Areas in which chemicals are underrepresented, termed diversity voids by (Pearlman & Smith 1998), lack relevant structural information for training models. Consequently, the underrepresented compounds become structural outliers for which the QSAR model is no longer instructive (Golbraikh 2000). Thus, chemical diversity and representativity should be considered during the collection of experimental data. While this is practiced in drug discovery to optimize sampling of a diverse chemical space for assay characterization (Bayada et al. 1999), this is not yet the case in toxicity testing which is more often driven by factors related to public health (e.g. use by vulnerable populations) and commerce (e.g. production volume) than chemistry.

For example, the comparatively poor prediction performance of QSAR models in the toxicogenomics project in Chapter 2 was due to numerous diversity voids in the chemical space (Figure 2.6A). The 127 compounds in the data set were selected to represent various hepatotoxic modes of action (e.g. phospholipidosis, glutathione depletion) rather than chemical diversity. As such, certain drug classes such as fibrates were overrepresented while others such as xanthines were underrepresented, resulting in an unevenly distributed chemical space with considerable diversity voids and activity cliffs (Sections 2.4 and 2.5). On average, the 127 compounds had low chemical similarities with their nearest neighbors (mean Tanimoto coefficient=0.65) and could not be reliably predicted from the relatively dissimilar neighbors. In contrast, the more biologically similar neighbors (mean Tanimoto coefficient=0.79) provided more accurate predictions.

To improve QSAR models, diversity voids may be filled in with additional chemical data points such that the voids are longer underrepresented. Such data set enrichment tries to span the largest chemical space with the fewest compounds in order to minimize data collection costs (Willett 2000). Data set enrichment may also resolve activity cliffs as additional data points with consistent SAR patterns dilute the activity outlier's contribution. In short, where QSAR modeling is expected, data sets should be designed with chemical diversity and representativity in mind.

5.3.2. Biological data limitations and future solutions

First, the bioassay data carried a certain degree of experimental error. Second, they were selected according to their availability instead of a rational basis. Because we had limited ourselves to bioassays that were generated under the same experimental conditions across all chemicals, only toxicogenomics and cytotoxicity assays from three experiments, TG-GATES, DrugMatrix and NCGC were used (Section 3.3.1). Of the chemicals with bioassays, toxicity labels for training models were extracted for as many chemicals as possible but might not necessarily relate to the bioassays by design. As such, the bioassays may not be useful predictors of the toxicity labels.

Chapter 3 showed that toxicogenomics assays, rich descriptions of general biological processes, were often better predictors of toxicity than cytotoxicity assays (Sections 3.4.2 and 3.5.6). In particular, cytotoxicity assays predicted mutagenicity with only 50% balanced accuracy. One possible reason may be the lack of biological relevance between cytotoxicity and mutagenicity which makes the case for rational selection of bioassays. In response, a framework for estimating toxicity centered on biological pathways has been proposed (Judson et al. 2011). Instead of predicting toxicity from all available bioassay data, only bioassays relevant to the pathway will be used. This will help to focus testing and modeling resources and improve overall model interpretation. Indeed, rational selection of bioassays by gene and protein functions increased prediction performance in several cases (Thomas et al. 2012).

Despite the array of bioassays available, important toxicological determinants such as toxicokinetics factors and exposure measures remain unaccounted for. Attempts to include bioavailability and metabolic clearance improved *in vitro-in vivo* extrapolation models (Rotroff et al. 2010, Wetmore et al. 2011). In light of these results, a future direction for CBRA may be to incorporate other toxicological factors in addition to the current use of chemical structures and bioassays.

5.3.3. Methodological limitations and future solutions

Further predictivity improvements may be achieved by more advanced machine learning techniques beyond the mainstream approaches used here. These advanced techniques exploit hidden data structures for learning and in doing so, arrive at more predictive models. Several examples include multi-task learning, which cross-learns from correlated toxicity activities (Caruana 1997, Zhang et al. 2013), and Bayesian hierarchical modeling, which can adjust for heterogeneous data with different degrees of error such that the different data sources are more effectively pooled for modeling (Liang et al. 2009). The latter recognizes that experimentally derived bioassays, being “noisier” than computed chemical descriptors, should be handled differently and allows the formulation of within-lab and between-lab errors.

Although this dissertation drew upon several methods besides machine learning classification (e.g. SOM clustering and network community detection), many more methods remain at our disposal. Exploratory analysis such as canonical correlation analysis may explain how two feature spaces (e.g. chemical and biological) are related. Regularized CCA was successfully used by Pauwels et al. to investigate how drug chemical structures relate to multiple ADR (Pauwels et al. 2011). The same technique may be applied to study how chemical structures relate to bioassays to generate key chemico-biological insights for subsequent examination.

Additional analysis, while not necessarily performed for prediction, may generate insight to support the interpretation of the prediction models. For example, the interpretation framework in Chapter 4 found a subset of chemical features associated with the SJS phenotype, knowledge that aided the development of a more parsimonious and interpretable QSAR model. Such feature selection, because of its parallels with genome-wide association studies (GWAS) (Moore et al. 2010, Touw et al. 2013), could be termed “Chemical-Wide Association Studies” (CWAS) (Low et al. 2013b), similar to other GWAS analogies such as Environmental-WAS (Patel et al. 2010) and Phenome-WAS) (Denny et al. 2010). Further development of CWAS could mirror the extensions made to GWAS, for example, to characterize fragment-fragment interactions akin to characterizing epistatic interactions (Cordell 2009, Winham et al. 2012).

Methodological challenges specific to each chapter include SJS being too rare an outcome for conclusive pharmacoepidemiology analysis. Its low incidence (one SJS case per 2-10 million patient-years, (Chan et al. 1990, Strom et al. 1991a)) meant that very few cases could be observed from our data set of 13 million patients, despite being extracted from the largest database of health insurance claims. Furthermore, the drugs of interest were rarely used and hence, provided a limited subpopulation that was underpowered for analysis (Section 4.6). In contrast, a more common ADR such as cardiotoxicity may be a better demonstration of the combined prowess of QSAR modeling and pharmacoepidemiology.

Confounder adjustment was limited to mainly demographic factors as only claims data were used. Additional data sources (clinical notes, laboratory reports) may provide a richer medical history for adjustment. Another major drawback relates to the way reference drugs were defined for training the SJS prediction models. Drugs should have been clinically defined by a gold standard instead of being statistically defined by a disproportionately high frequency of SJS spontaneous reports (Section 4.3.1).

Temporal analysis was not performed due to the lack of temporal data. Benefits in terms of earlier SJS prediction could be demonstrated through a prospective validation which sets aside new data for validating models derived from older data. Examples of effective use of prospective validation include (Cami et al. 2011) which showed that simultaneous modeling of multiple ADR and drugs resulted in earlier detection of ADR and (Baker 2010) which showed that earlier biomedical literature foretold later disease-chemical-protein associations.

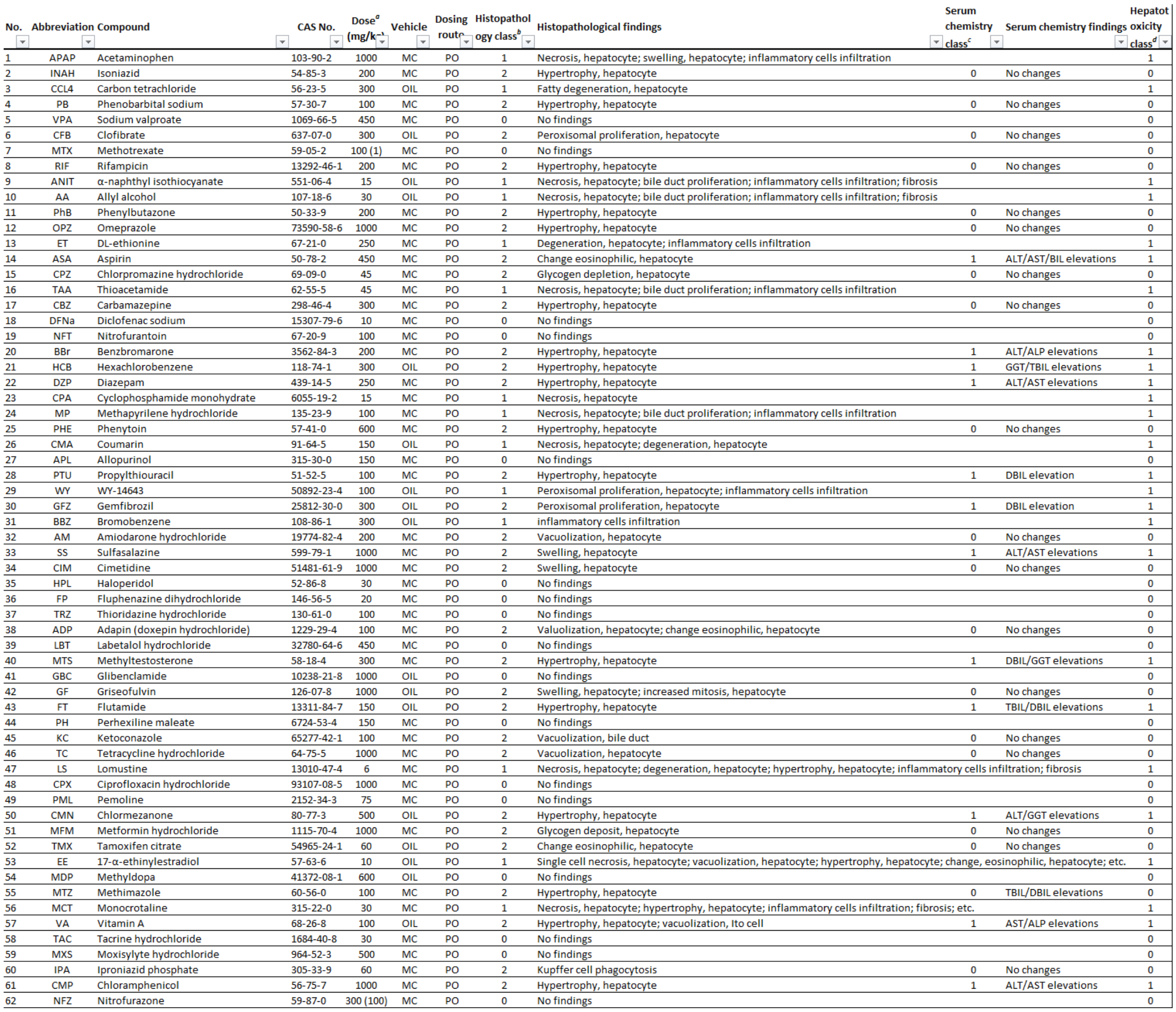
5.4. Epilogue

In closing, this work demonstrated improved predictivity and interpretation of toxicity models through several integrative approaches, drawing data and methods from cheminformatics, bioinformatics and pharmacoepidemiology. The work presented two of many possible integrative approaches that could benefit the study of toxicology. One can envision further gains through tighter and broader integration of the various disciplines. Tighter integration calls for greater familiarity of the data and techniques across disciplines to facilitate interaction. Broader integration adopts an inclusive outlook to draw from as many relevant disciplines as possible. The ingredients for such multi-disciplinary efforts are unlikely to occur organically and will require deliberate efforts to foster a collaborative environment. I hope that this work has demonstrated the practical benefits of integrative methods and will motivate further research to develop more accurate and insightful toxicity prediction models towards safer chemicals and healthier lives.

APPENDIX 1: SUPPLEMENTAL TABLES

Supplemental tables for Chapter 2 (also available online at doi:10.1021/tx200148a)

Table A1.2.I. List of compounds and their hepatotoxicities



No.	Abbreviation	Compound	CAS No.	Dose[a] (mg/kg)	Vehicle	Dosing route	Histopathology class[b]	Histopathological findings	Serum chemistry class[c]	Serum chemistry findings	Hepatotoxicity class[d]
1	APAP	Acetaminophen	103-90-2	1000	MC	PO	1	Necrosis, hepatocyte; swelling, hepatocyte; inflammatory cells infiltration			1
2	INAH	Isoniazid	54-85-3	200	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
3	CCL4	Carbon tetrachloride	56-23-5	300	OIL	PO	1	Fatty degeneration, hepatocyte			1
4	PB	Phenobarbital sodium	57-30-7	100	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
5	VPA	Sodium valproate	1069-66-5	450	MC	PO	0	No findings			0
6	CFB	Clofibrate	637-07-0	300	OIL	PO	2	Peroxisomal proliferation, hepatocyte	0	No changes	0
7	MTX	Methotrexate	59-05-2	100 (1)	MC	PO	0	No findings			0
8	RIF	Rifampicin	13292-46-1	200	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
9	ANIT	α-naphthyl isothiocyanate	551-06-4	15	OIL	PO	1	Necrosis, hepatocyte; bile duct proliferation; inflammatory cells infiltration; fibrosis			1
10	AA	Allyl alcohol	107-18-6	30	OIL	PO	1	Necrosis, hepatocyte; bile duct proliferation; inflammatory cells infiltration; fibrosis			1
11	PhB	Phenylbutazone	50-33-9	200	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
12	OPZ	Omeprazole	73590-58-6	1000	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
13	ET	DL-ethionine	67-21-0	250	MC	PO	1	Degeneration, hepatocyte; inflammatory cells infiltration			1
14	ASA	Aspirin	50-78-2	450	MC	PO	2	Change eosinophilic, hepatocyte	1	ALT/AST/BIL elevations	1
15	CPZ	Chlorpromazine hydrochloride	69-09-0	45	MC	PO	2	Glycogen depletion, hepatocyte	0	No changes	0
16	TAA	Thioacetamide	62-55-5	45	MC	PO	1	Necrosis, hepatocyte; bile duct proliferation; inflammatory cells infiltration			1
17	CBZ	Carbamazepine	298-46-4	300	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
18	DFNa	Diclofenac sodium	15307-79-6	10	MC	PO	0	No findings			0
19	NFT	Nitrofurantoin	67-20-9	100	MC	PO	0	No findings			0
20	BBr	Benzbromarone	3562-84-3	200	MC	PO	2	Hypertrophy, hepatocyte	1	ALT/ALP elevations	1
21	HCB	Hexachlorobenzene	118-74-1	300	OIL	PO	2	Hypertrophy, hepatocyte	1	GGT/TBIL elevations	1
22	DZP	Diazepam	439-14-5	250	MC	PO	2	Hypertrophy, hepatocyte	1	ALT/AST elevations	1
23	CPA	Cyclophosphamide monohydrate	6055-19-2	15	MC	PO	1	Necrosis, hepatocyte			1
24	MP	Methapyrilene hydrochloride	135-23-9	100	MC	PO	1	Necrosis, hepatocyte; bile duct proliferation; inflammatory cells infiltration			1
25	PHE	Phenytoin	57-41-0	600	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
26	CMA	Coumarin	91-64-5	150	OIL	PO	1	Necrosis, hepatocyte; degeneration, hepatocyte			1
27	APL	Allopurinol	315-30-0	150	MC	PO	0	No findings			0
28	PTU	Propylthiouracil	51-52-5	100	MC	PO	2	Hypertrophy, hepatocyte	1	DBIL elevation	1
29	WY	WY-14643	50892-23-4	100	OIL	PO	1	Peroxisomal proliferation, hepatocyte; inflammatory cells infiltration			1
30	GFZ	Gemfibrozil	25812-30-0	300	OIL	PO	2	Peroxisomal proliferation, hepatocyte	1	DBIL elevation	1
31	BBZ	Bromobenzene	108-86-1	300	OIL	PO	1	inflammatory cells infiltration			1
32	AM	Amiodarone hydrochloride	19774-82-4	200	MC	PO	2	Vacuolization, hepatocyte	0	No changes	0
33	SS	Sulfasalazine	599-79-1	1000	MC	PO	2	Swelling, hepatocyte	1	ALT/AST elevations	1
34	CIM	Cimetidine	51481-61-9	1000	MC	PO	2	Swelling, hepatocyte	0	No changes	0
35	HPL	Haloperidol	52-86-8	30	MC	PO	0	No findings			0
36	FP	Fluphenazine dihydrochloride	146-56-5	20	MC	PO	0	No findings			0
37	TRZ	Thioridazine hydrochloride	130-61-0	100	MC	PO	0	No findings			0
38	ADP	Adapin (doxepin hydrochloride)	1229-29-4	100	MC	PO	2	Valuolization, hepatocyte; change eosinophilic, hepatocyte	0	No changes	0
39	LBT	Labetalol hydrochloride	32780-64-6	450	MC	PO	0	No findings			0
40	MTS	Methyltestosterone	58-18-4	300	MC	PO	2	Hypertrophy, hepatocyte	1	DBIL/GGT elevations	1
41	GBC	Glibenclamide	10238-21-8	1000	OIL	PO	0	No findings			0
42	GF	Griseofulvin	126-07-8	1000	OIL	PO	2	Swelling, hepatocyte; increased mitosis, hepatocyte	0	No changes	0
43	FT	Flutamide	13311-84-7	150	OIL	PO	2	Hypertrophy, hepatocyte	1	TBIL/DBIL elevations	1
44	PH	Perhexiline maleate	6724-53-4	150	MC	PO	0	No findings			0
45	KC	Ketoconazole	65277-42-1	100	MC	PO	2	Vacuolization, bile duct	0	No changes	0
46	TC	Tetracycline hydrochloride	64-75-5	1000	MC	PO	2	Vacuolization, hepatocyte	0	No changes	0
47	LS	Lomustine	13010-47-4	6	MC	PO	1	Necrosis, hepatocyte; degeneration, hepatocyte; hypertrophy, hepatocyte; inflammatory cells infiltration; fibrosis			1
48	CPX	Ciprofloxacin hydrochloride	93107-08-5	1000	MC	PO	0	No findings			0
49	PML	Pemoline	2152-34-3	75	MC	PO	0	No findings			0
50	CMN	Chlormezanone	80-77-3	500	OIL	PO	2	Hypertrophy, hepatocyte	1	ALT/GGT elevations	1
51	MFM	Metformin hydrochloride	1115-70-4	1000	MC	PO	2	Glycogen deposit, hepatocyte	0	No changes	0
52	TMX	Tamoxifen citrate	54965-24-1	60	OIL	PO	2	Change eosinophilic, hepatocyte	0	No changes	0
53	EE	17-α-ethinylestradiol	57-63-6	10	OIL	PO	1	Single cell necrosis, hepatocyte; vacuolization, hepatocyte; hypertrophy, hepatocyte; change, eosinophilic, hepatocyte; etc.			1
54	MDP	Methyldopa	41372-08-1	600	OIL	PO	0	No findings			0
55	MTZ	Methimazole	60-56-0	100	MC	PO	2	Hypertrophy, hepatocyte	0	TBIL/DBIL elevations	0
56	MCT	Monocrotaline	315-22-0	30	MC	PO	1	Necrosis, hepatocyte; hypertrophy, hepatocyte; inflammatory cells infiltration; fibrosis; etc.			1
57	VA	Vitamin A	68-26-8	100	OIL	PO	2	Hypertrophy, hepatocyte; vacuolization, Ito cell	1	AST/ALP elevations	1
58	TAC	Tacrine hydrochloride	1684-40-8	30	MC	PO	0	No findings			0
59	MXS	Moxisylyte hydrochloride	964-52-3	500	MC	PO	0	No findings			0
60	IPA	Iproniazid phosphate	305-33-9	60	MC	PO	2	Kupffer cell phagocytosis	0	No changes	0
61	CMP	Chloramphenicol	56-75-7	1000	MC	PO	2	Hypertrophy, hepatocyte	1	ALT/AST elevations	1
62	NFZ	Nitrofurazone	59-87-0	300 (100)	MC	PO	0	No findings			0

[a] Doses in single dosing and repeated dosing studies are identical unless denoted in parenthesis as repeated dose.

[b] Histopathology classes: 0 = no hepatotoxic findings, 1 = hepatotoxic findings, or 2 = other findings

[c] Serum chemistry classes: 0 = no changes in biochemical markers or 1 = changes in biochemical markers

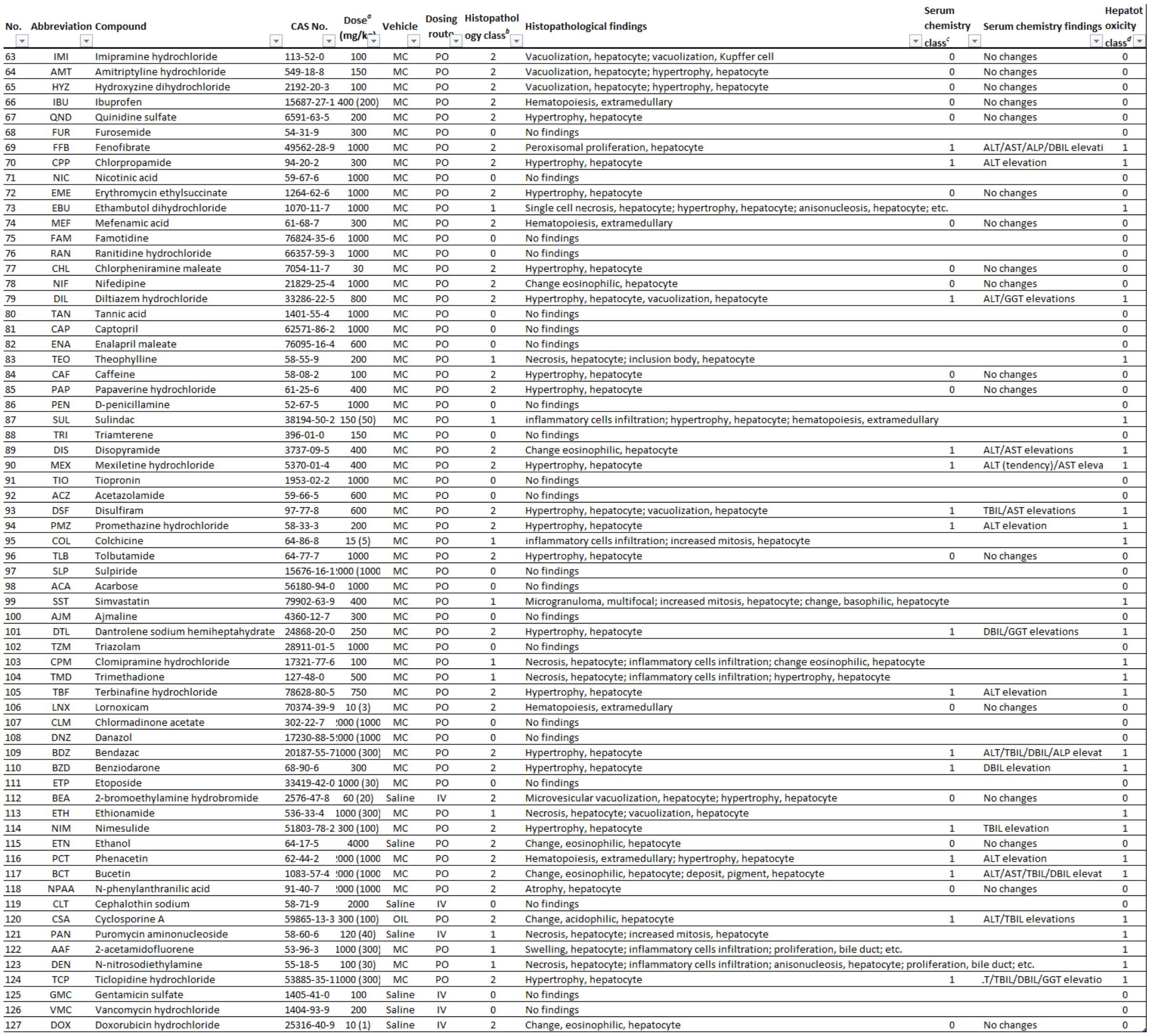
[d] Heptotoxicity classes: 0 = non-hepatotoxic or 1 = hepatotoxic

Histopathology class is denoted "0" when no histopathology findings are observed, "1" when hepatocell necrosis, degeneration, or inflammation are present, and "2" for all other findings.

The determination of hepatotoxicity was further augmented by the statistically significant elevation of the serum chemistry biomarkers: alanine aminotransferase (ALT), asparate aminotransferase (AST), alkaline phosphatase (ALP), total bilirubin (TBIL), direct bilirubin (DBIL), and gamma-glutamyl transpeptidase (GGT).

A compound is hepatotoxic (1) if histopathology alone showed hepatotoxicity (histopathology class=1), or it has other histopathology findings (histopathology class=2) and elevated serum chemistry (blood chemistry class=1). A compound is non-hepatotoxic(0) if no histopathology findings were observed (histopathology class=0), or it has other histopathology findings (histopathology class=2) but normal serum chemistry (blood chemistry class=1).

Table A1.2.I (continued). List of compounds and their hepatotoxicities



No.	Abbreviation	Compound	CAS No.	Dose[a] (mg/k	Vehicle	Dosing route	Histopathology class[b]	Histopathological findings	Serum chemistry class[c]	Serum chemistry findings	Hepatotoxicity class[d]
63	IMI	Imipramine hydrochloride	113-52-0	100	MC	PO	2	Vacuolization, hepatocyte; vacuolization, Kupffer cell	0	No changes	0
64	AMT	Amitriptyline hydrochloride	549-18-8	150	MC	PO	2	Vacuolization, hepatocyte; hypertrophy, hepatocyte	0	No changes	0
65	HYZ	Hydroxyzine dihydrochloride	2192-20-3	100	MC	PO	2	Vacuolization, hepatocyte; hypertrophy, hepatocyte	0	No changes	0
66	IBU	Ibuprofen	15687-27-1	400 (200)	MC	PO	2	Hematopoiesis, extramedullary	0	No changes	0
67	QND	Quinidine sulfate	6591-63-5	200	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
68	FUR	Furosemide	54-31-9	300	MC	PO	0	No findings			0
69	FFB	Fenofibrate	49562-28-9	1000	MC	PO	2	Peroxisomal proliferation, hepatocyte	1	ALT/AST/ALP/DBIL elevati	1
70	CPP	Chlorpropamide	94-20-2	300	MC	PO	2	Hypertrophy, hepatocyte	1	ALT elevation	1
71	NIC	Nicotinic acid	59-67-6	1000	MC	PO	0	No findings			0
72	EME	Erythromycin ethylsuccinate	1264-62-6	1000	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
73	EBU	Ethambutol dihydrochloride	1070-11-7	1000	MC	PO	1	Single cell necrosis, hepatocyte; hypertrophy, hepatocyte; anisonucleosis, hepatocyte; etc.			1
74	MEF	Mefenamic acid	61-68-7	300	MC	PO	2	Hematopoiesis, extramedullary	0	No changes	0
75	FAM	Famotidine	76824-35-6	1000	MC	PO	0	No findings			0
76	RAN	Ranitidine hydrochloride	66357-59-3	1000	MC	PO	0	No findings			0
77	CHL	Chlorpheniramine maleate	7054-11-7	30	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
78	NIF	Nifedipine	21829-25-4	1000	MC	PO	2	Change eosinophilic, hepatocyte	0	No changes	0
79	DIL	Diltiazem hydrochloride	33286-22-5	800	MC	PO	2	Hypertrophy, hepatocyte, vacuolization, hepatocyte	1	ALT/GGT elevations	1
80	TAN	Tannic acid	1401-55-4	1000	MC	PO	0	No findings			0
81	CAP	Captopril	62571-86-2	1000	MC	PO	0	No findings			0
82	ENA	Enalapril maleate	76095-16-4	600	MC	PO	0	No findings			0
83	TEO	Theophylline	58-55-9	200	MC	PO	1	Necrosis, hepatocyte; inclusion body, hepatocyte			1
84	CAF	Caffeine	58-08-2	100	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
85	PAP	Papaverine hydrochloride	61-25-6	400	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
86	PEN	D-penicillamine	52-67-5	1000	MC	PO	0	No findings			0
87	SUL	Sulindac	38194-50-2	150 (50)	MC	PO	1	inflammatory cells infiltration; hypertrophy, hepatocyte; hematopoiesis, extramedullary			1
88	TRI	Triamterene	396-01-0	150	MC	PO	0	No findings			0
89	DIS	Disopyramide	3737-09-5	400	MC	PO	2	Change eosinophilic, hepatocyte	1	ALT/AST elevations	1
90	MEX	Mexiletine hydrochloride	5370-01-4	400	MC	PO	2	Hypertrophy, hepatocyte	1	ALT (tendency)/AST eleva	1
91	TIO	Tiopronin	1953-02-2	1000	MC	PO	0	No findings			0
92	ACZ	Acetazolamide	59-66-5	600	MC	PO	0	No findings			0
93	DSF	Disulfiram	97-77-8	600	MC	PO	2	Hypertrophy, hepatocyte; vacuolization, hepatocyte	1	TBIL/AST elevations	1
94	PMZ	Promethazine hydrochloride	58-33-3	200	MC	PO	2	Hypertrophy, hepatocyte	1	ALT elevation	1
95	COL	Colchicine	64-86-8	15 (5)	MC	PO	1	inflammatory cells infiltration; increased mitosis, hepatocyte			1
96	TLB	Tolbutamide	64-77-7	1000	MC	PO	2	Hypertrophy, hepatocyte	0	No changes	0
97	SLP	Sulpiride	15676-16-1	!000 (1000	MC	PO	0	No findings			0
98	ACA	Acarbose	56180-94-0	1000	MC	PO	0	No findings			0
99	SST	Simvastatin	79902-63-9	400	MC	PO	1	Microgranuloma, multifocal; increased mitosis, hepatocyte; change, basophilic, hepatocyte			1
100	AJM	Ajmaline	4360-12-7	300	MC	PO	0	No findings			0
101	DTL	Dantrolene sodium hemiheptahydrate	24868-20-0	250	MC	PO	2	Hypertrophy, hepatocyte	1	DBIL/GGT elevations	1
102	TZM	Triazolam	28911-01-5	1000	MC	PO	0	No findings			0
103	CPM	Clomipramine hydrochloride	17321-77-6	100	MC	PO	1	Necrosis, hepatocyte; inflammatory cells infiltration; change eosinophilic, hepatocyte			1
104	TMD	Trimethadione	127-48-0	500	MC	PO	1	Necrosis, hepatocyte; inflammatory cells infiltration; hypertrophy, hepatocyte			1
105	TBF	Terbinafine hydrochloride	78628-80-5	750	MC	PO	2	Hypertrophy, hepatocyte	1	ALT elevation	1
106	LNX	Lornoxicam	70374-39-9	10 (3)	MC	PO	2	Hematopoiesis, extramedullary	0	No changes	0
107	CLM	Chlormadinone acetate	302-22-7	!000 (1000	MC	PO	0	No findings			0
108	DNZ	Danazol	17230-88-5	!000 (1000	MC	PO	0	No findings			0
109	BDZ	Bendazac	20187-55-7	1000 (300)	MC	PO	2	Hypertrophy, hepatocyte	1	ALT/TBIL/DBIL/ALP elevat	1
110	BZD	Benziodarone	68-90-6	300	MC	PO	2	Hypertrophy, hepatocyte	1	DBIL elevation	1
111	ETP	Etoposide	33419-42-0	1000 (30)	MC	PO	0	No findings			0
112	BEA	2-bromoethylamine hydrobromide	2576-47-8	60 (20)	Saline	IV	2	Microvesicular vacuolization, hepatocyte; hypertrophy, hepatocyte	0	No changes	0
113	ETH	Ethionamide	536-33-4	1000 (300)	MC	PO	1	Necrosis, hepatocyte; vacuolization, hepatocyte			1
114	NIM	Nimesulide	51803-78-2	300 (100)	MC	PO	2	Hypertrophy, hepatocyte	1	TBIL elevation	1
115	ETN	Ethanol	64-17-5	4000	Saline	PO	2	Change, eosinophilic, hepatocyte	0	No changes	0
116	PCT	Phenacetin	62-44-2	!000 (1000	MC	PO	2	Hematopoiesis, extramedullary; hypertrophy, hepatocyte	1	ALT elevation	1
117	BCT	Bucetin	1083-57-4	!000 (1000	MC	PO	2	Change, eosinophilic, hepatocyte; deposit, pigment, hepatocyte	1	ALT/AST/TBIL/DBIL elevat	1
118	NPAA	N-phenylanthranilic acid	91-40-7	!000 (1000	MC	PO	2	Atrophy, hepatocyte	0	No changes	0
119	CLT	Cephalothin sodium	58-71-9	2000	Saline	IV	0	No findings			0
120	CSA	Cyclosporine A	59865-13-3	300 (100)	OIL	PO	2	Change, acidophilic, hepatocyte	1	ALT/TBIL elevations	1
121	PAN	Puromycin aminonucleoside	58-60-6	120 (40)	Saline	IV	1	Necrosis, hepatocyte; increased mitosis, hepatocyte			1
122	AAF	2-acetamidofluorene	53-96-3	1000 (300)	MC	PO	1	Swelling, hepatocyte; inflammatory cells infiltration; proliferation, bile duct; etc.			1
123	DEN	N-nitrosodiethylamine	55-18-5	100 (30)	MC	PO	1	Necrosis, hepatocyte; inflammatory cells infiltration; anisonucleosis, hepatocyte; proliferation, bile duct; etc.			1
124	TCP	Ticlopidine hydrochloride	53885-35-1	1000 (300)	MC	PO	2	Hypertrophy, hepatocyte	1	.T/TBIL/DBIL/GGT elevatio	1
125	GMC	Gentamicin sulfate	1405-41-0	100	Saline	IV	0	No findings			0
126	VMC	Vancomycin hydrochloride	1404-93-9	200	Saline	IV	0	No findings			0
127	DOX	Doxorubicin hydrochloride	25316-40-9	10 (1)	Saline	IV	2	Change, eosinophilic, hepatocyte	0	No changes	0

[a] Doses in single dosing and repeated dosing studies are identical unless denoted in parenthesis as repeated dose.

[b] Histopathology classes: 0 = no hepatotoxic findings, 1 = hepatotoxic findings, or 2 = other findings

[c] Serum chemistry classes: 0 = no changes in biochemical markers or 1 = changes in biochemical markers

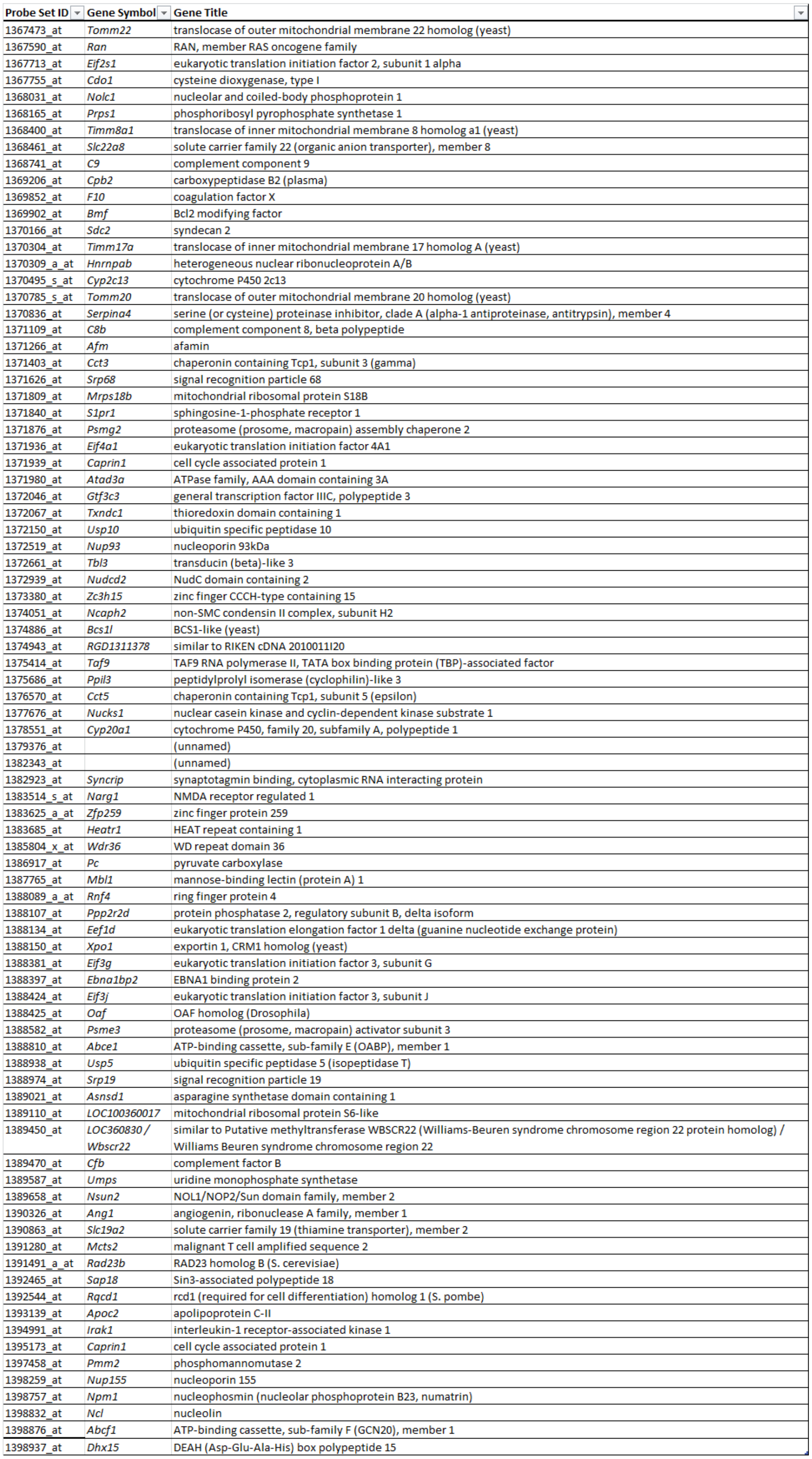
[d] Heptotoxicity classes: 0 = non-hepatotoxic or 1 = hepatotoxic

Histopathology class is denoted "0" when no histopathology findings are observed, "1" when hepatocell necrosis, degeneration, or inflammation are present, and "2" for all other findings.

The determination of hepatotoxicity was further augmented by the statistically significant elevation of the serum chemistry biomarkers: alanine aminotransferase (ALT), asparate aminotransferase (AST), alkaline phosphatase (ALP), total bilirubin (TBIL), direct bilirubin (DBIL), and gamma-glutamyl transpeptidase (GGT).

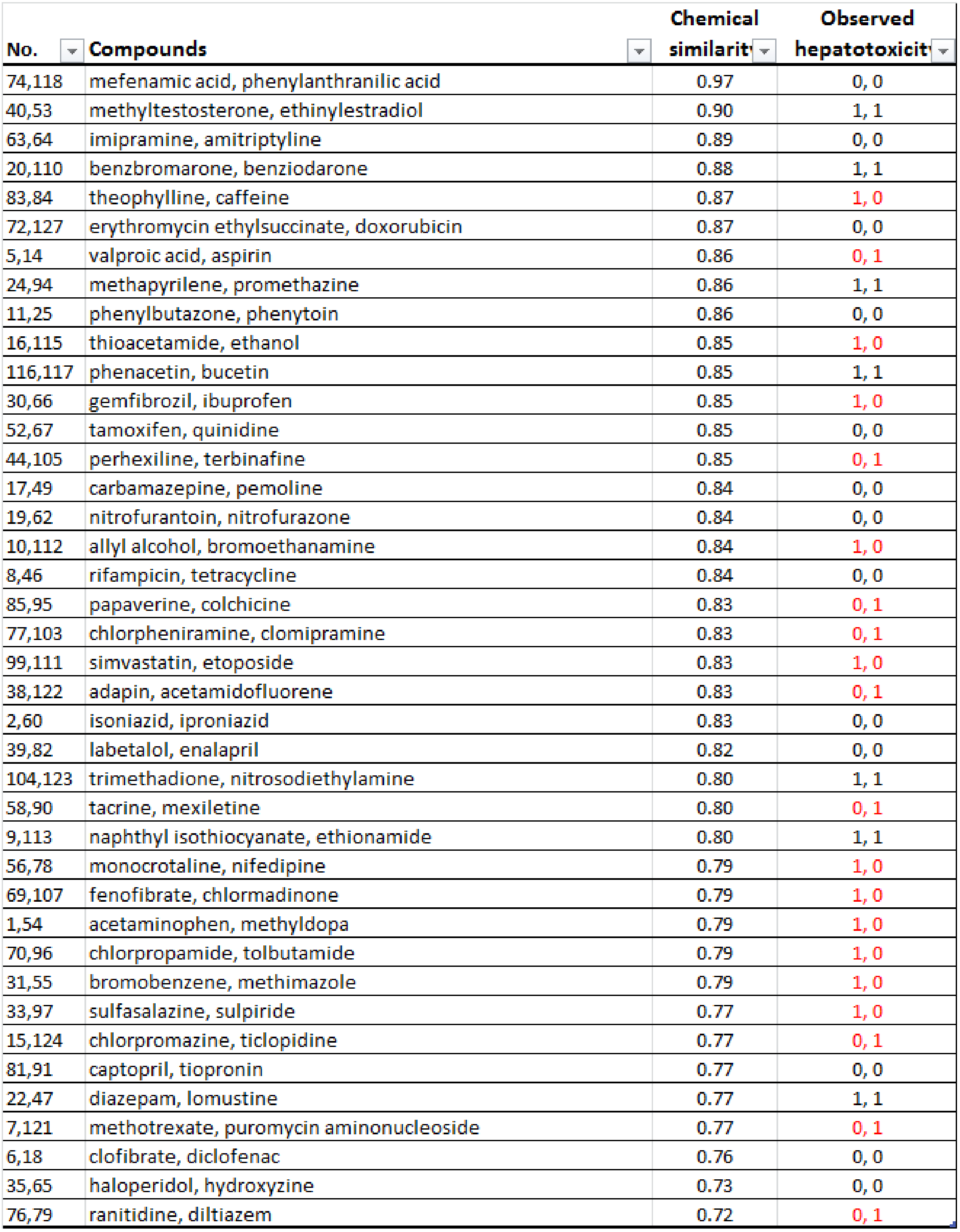
A compound is hepatotoxic (1) if histopathology alone showed hepatotoxicity (histopathology class=1), or it has other histopathology findings (histopathology class=2) and elevated serum chemistry (blood chemistry class=1).A compound is non-hepatotoxic(0) if no histopathology findings were observed (histopathology class=0), or it has other histopathology findings (histopathology class=2) but normal serum chemistry (blood chemistry class=1).

Table A1.2.II. List of 85 predictive genes



Probe Set ID	Gene Symbol	Gene Title
1367473_at	*Tomm22*	translocase of outer mitochondrial membrane 22 homolog (yeast)
1367590_at	*Ran*	RAN, member RAS oncogene family
1367713_at	*Eif2s1*	eukaryotic translation initiation factor 2, subunit 1 alpha
1367755_at	*Cdo1*	cysteine dioxygenase, type I
1368031_at	*Nolc1*	nucleolar and coiled-body phosphoprotein 1
1368165_at	*Prps1*	phosphoribosyl pyrophosphate synthetase 1
1368400_at	*Timm8a1*	translocase of inner mitochondrial membrane 8 homolog a1 (yeast)
1368461_at	*Slc22a8*	solute carrier family 22 (organic anion transporter), member 8
1368741_at	*C9*	complement component 9
1369206_at	*Cpb2*	carboxypeptidase B2 (plasma)
1369852_at	*F10*	coagulation factor X
1369902_at	*Bmf*	Bcl2 modifying factor
1370166_at	*Sdc2*	syndecan 2
1370304_at	*Timm17a*	translocase of inner mitochondrial membrane 17 homolog A (yeast)
1370309_a_at	*Hnrnpab*	heterogeneous nuclear ribonucleoprotein A/B
1370495_s_at	*Cyp2c13*	cytochrome P450 2c13
1370785_s_at	*Tomm20*	translocase of outer mitochondrial membrane 20 homolog (yeast)
1370836_at	*Serpina4*	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 4
1371109_at	*C8b*	complement component 8, beta polypeptide
1371266_at	*Afm*	afamin
1371403_at	*Cct3*	chaperonin containing Tcp1, subunit 3 (gamma)
1371626_at	*Srp68*	signal recognition particle 68
1371809_at	*Mrps18b*	mitochondrial ribosomal protein S18B
1371840_at	*S1pr1*	sphingosine-1-phosphate receptor 1
1371876_at	*Psmg2*	proteasome (prosome, macropain) assembly chaperone 2
1371936_at	*Eif4a1*	eukaryotic translation initiation factor 4A1
1371939_at	*Caprin1*	cell cycle associated protein 1
1371980_at	*Atad3a*	ATPase family, AAA domain containing 3A
1372046_at	*Gtf3c3*	general transcription factor IIIC, polypeptide 3
1372067_at	*Txndc1*	thioredoxin domain containing 1
1372150_at	*Usp10*	ubiquitin specific peptidase 10
1372519_at	*Nup93*	nucleoporin 93kDa
1372661_at	*Tbl3*	transducin (beta)-like 3
1372939_at	*Nudcd2*	NudC domain containing 2
1373380_at	*Zc3h15*	zinc finger CCCH-type containing 15
1374051_at	*Ncaph2*	non-SMC condensin II complex, subunit H2
1374886_at	*Bcs1l*	BCS1-like (yeast)
1374943_at	*RGD1311378*	similar to RIKEN cDNA 2010011I20
1375414_at	*Taf9*	TAF9 RNA polymerase II, TATA box binding protein (TBP)-associated factor
1375686_at	*Ppil3*	peptidylprolyl isomerase (cyclophilin)-like 3
1376570_at	*Cct5*	chaperonin containing Tcp1, subunit 5 (epsilon)
1377676_at	*Nucks1*	nuclear casein kinase and cyclin-dependent kinase substrate 1
1378551_at	*Cyp20a1*	cytochrome P450, family 20, subfamily A, polypeptide 1
1379376_at		(unnamed)
1382343_at		(unnamed)
1382923_at	*Syncrip*	synaptotagmin binding, cytoplasmic RNA interacting protein
1383514_s_at	*Narg1*	NMDA receptor regulated 1
1383625_a_at	*Zfp259*	zinc finger protein 259
1383685_at	*Heatr1*	HEAT repeat containing 1
1385804_x_at	*Wdr36*	WD repeat domain 36
1386917_at	*Pc*	pyruvate carboxylase
1387765_at	*Mbl1*	mannose-binding lectin (protein A) 1
1388089_a_at	*Rnf4*	ring finger protein 4
1388107_at	*Ppp2r2d*	protein phosphatase 2, regulatory subunit B, delta isoform
1388134_at	*Eef1d*	eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein)
1388150_at	*Xpo1*	exportin 1, CRM1 homolog (yeast)
1388381_at	*Eif3g*	eukaryotic translation initiation factor 3, subunit G
1388397_at	*Ebna1bp2*	EBNA1 binding protein 2
1388424_at	*Eif3j*	eukaryotic translation initiation factor 3, subunit J
1388425_at	*Oaf*	OAF homolog (Drosophila)
1388582_at	*Psme3*	proteasome (prosome, macropain) activator subunit 3
1388810_at	*Abce1*	ATP-binding cassette, sub-family E (OABP), member 1
1388938_at	*Usp5*	ubiquitin specific peptidase 5 (isopeptidase T)
1388974_at	*Srp19*	signal recognition particle 19
1389021_at	*Asnsd1*	asparagine synthetase domain containing 1
1389110_at	*LOC100360017*	mitochondrial ribosomal protein S6-like
1389450_at	*LOC360830 / Wbscr22*	similar to Putative methyltransferase WBSCR22 (Williams-Beuren syndrome chromosome region 22 protein homolog) / Williams Beuren syndrome chromosome region 22
1389470_at	*Cfb*	complement factor B
1389587_at	*Umps*	uridine monophosphate synthetase
1389658_at	*Nsun2*	NOL1/NOP2/Sun domain family, member 2
1390326_at	*Ang1*	angiogenin, ribonuclease A family, member 1
1390863_at	*Slc19a2*	solute carrier family 19 (thiamine transporter), member 2
1391280_at	*Mcts2*	malignant T cell amplified sequence 2
1391491_a_at	*Rad23b*	RAD23 homolog B (S. cerevisiae)
1392465_at	*Sap18*	Sin3-associated polypeptide 18
1392544_at	*Rqcd1*	rcd1 (required for cell differentiation) homolog 1 (S. pombe)
1393139_at	*Apoc2*	apolipoprotein C-II
1394991_at	*Irak1*	interleukin-1 receptor-associated kinase 1
1395173_at	*Caprin1*	cell cycle associated protein 1
1397458_at	*Pmm2*	phosphomannomutase 2
1398259_at	*Nup155*	nucleoporin 155
1398757_at	*Npm1*	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
1398832_at	*Ncl*	nucleolin
1398876_at	*Abcf1*	ATP-binding cassette, sub-family F (GCN20), member 1
1398937_at	*Dhx15*	DEAH (Asp-Glu-Ala-His) box polypeptide 15

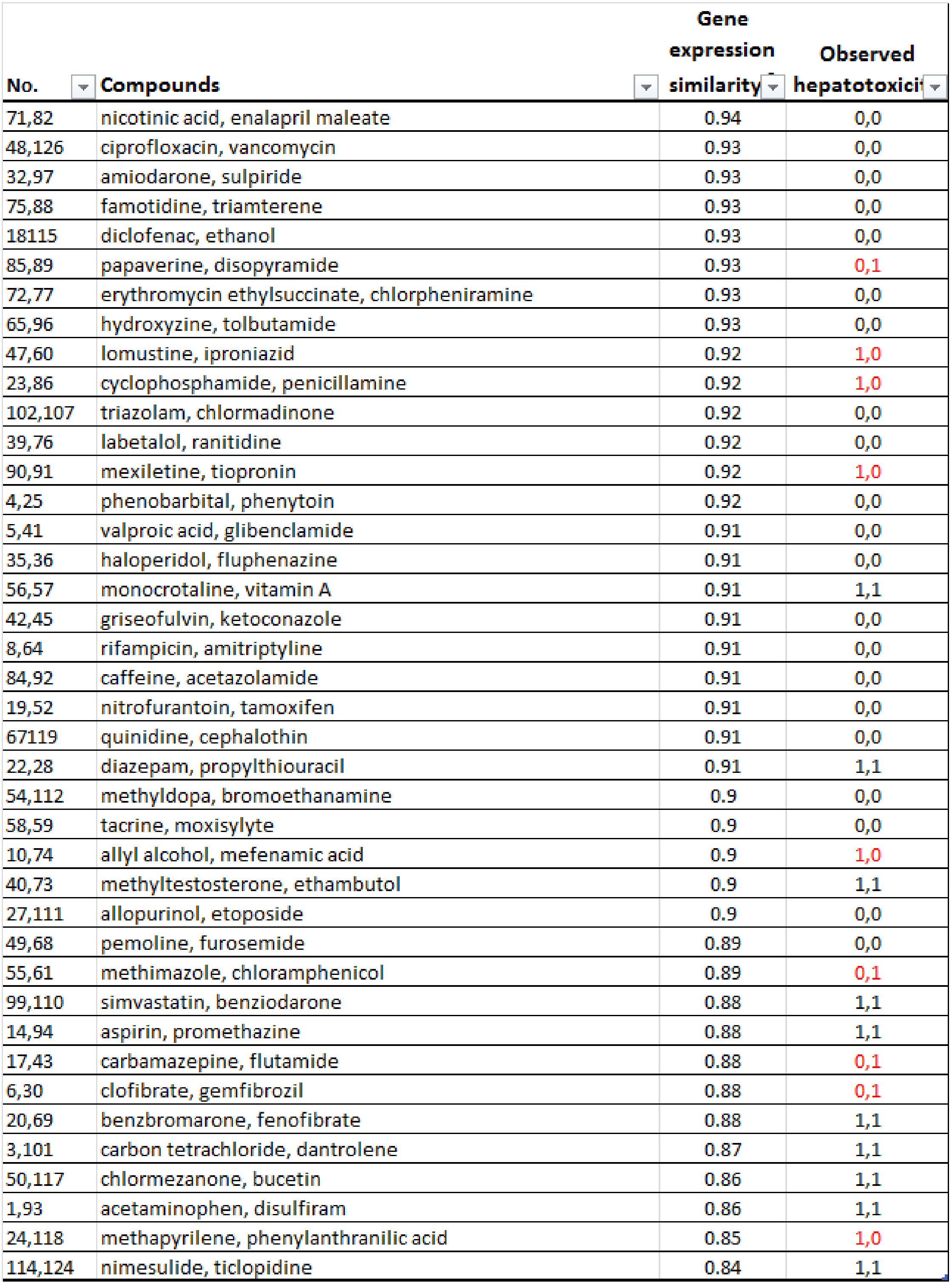
Table A1.2.IIIa. List of 40 chemically closest compound pairs and their hepatotoxicities. Twenty pairs (50%) have opposite toxicities despite their chemical similarity (0=dissimilar, 1=identical)[a]



No.	Compounds	Chemical similarit	Observed hepatotoxicit
74,118	mefenamic acid, phenylanthranilic acid	0.97	0, 0
40,53	methyltestosterone, ethinylestradiol	0.90	1, 1
63,64	imipramine, amitriptyline	0.89	0, 0
20,110	benzbromarone, benziodarone	0.88	1, 1
83,84	theophylline, caffeine	0.87	1, 0
72,127	erythromycin ethylsuccinate, doxorubicin	0.87	0, 0
5,14	valproic acid, aspirin	0.86	0, 1
24,94	methapyrilene, promethazine	0.86	1, 1
11,25	phenylbutazone, phenytoin	0.86	0, 0
16,115	thioacetamide, ethanol	0.85	1, 0
116,117	phenacetin, bucetin	0.85	1, 1
30,66	gemfibrozil, ibuprofen	0.85	1, 0
52,67	tamoxifen, quinidine	0.85	0, 0
44,105	perhexiline, terbinafine	0.85	0, 1
17,49	carbamazepine, pemoline	0.84	0, 0
19,62	nitrofurantoin, nitrofurazone	0.84	0, 0
10,112	allyl alcohol, bromoethanamine	0.84	1, 0
8,46	rifampicin, tetracycline	0.84	0, 0
85,95	papaverine, colchicine	0.83	0, 1
77,103	chlorpheniramine, clomipramine	0.83	0, 1
99,111	simvastatin, etoposide	0.83	1, 0
38,122	adapin, acetamidofluorene	0.83	0, 1
2,60	isoniazid, iproniazid	0.83	0, 0
39,82	labetalol, enalapril	0.82	0, 0
104,123	trimethadione, nitrosodiethylamine	0.80	1, 1
58,90	tacrine, mexiletine	0.80	0, 1
9,113	naphthyl isothiocyanate, ethionamide	0.80	1, 1
56,78	monocrotaline, nifedipine	0.79	1, 0
69,107	fenofibrate, chlormadinone	0.79	1, 0
1,54	acetaminophen, methyldopa	0.79	1, 0
70,96	chlorpropamide, tolbutamide	0.79	1, 0
31,55	bromobenzene, methimazole	0.79	1, 0
33,97	sulfasalazine, sulpiride	0.77	1, 0
15,124	chlorpromazine, ticlopidine	0.77	0, 1
81,91	captopril, tiopronin	0.77	0, 0
22,47	diazepam, lomustine	0.77	1, 1
7,121	methotrexate, puromycin aminonucleoside	0.77	0, 1
6,18	clofibrate, diclofenac	0.76	0, 0
35,65	haloperidol, hydroxyzine	0.73	0, 0
76,79	ranitidine, diltiazem	0.72	0, 1

[a] Similarity = 1 - normalized pairwise Euclidean distance.
Identical pairs have a similarity of 1; completely different pairs have zero similarity.

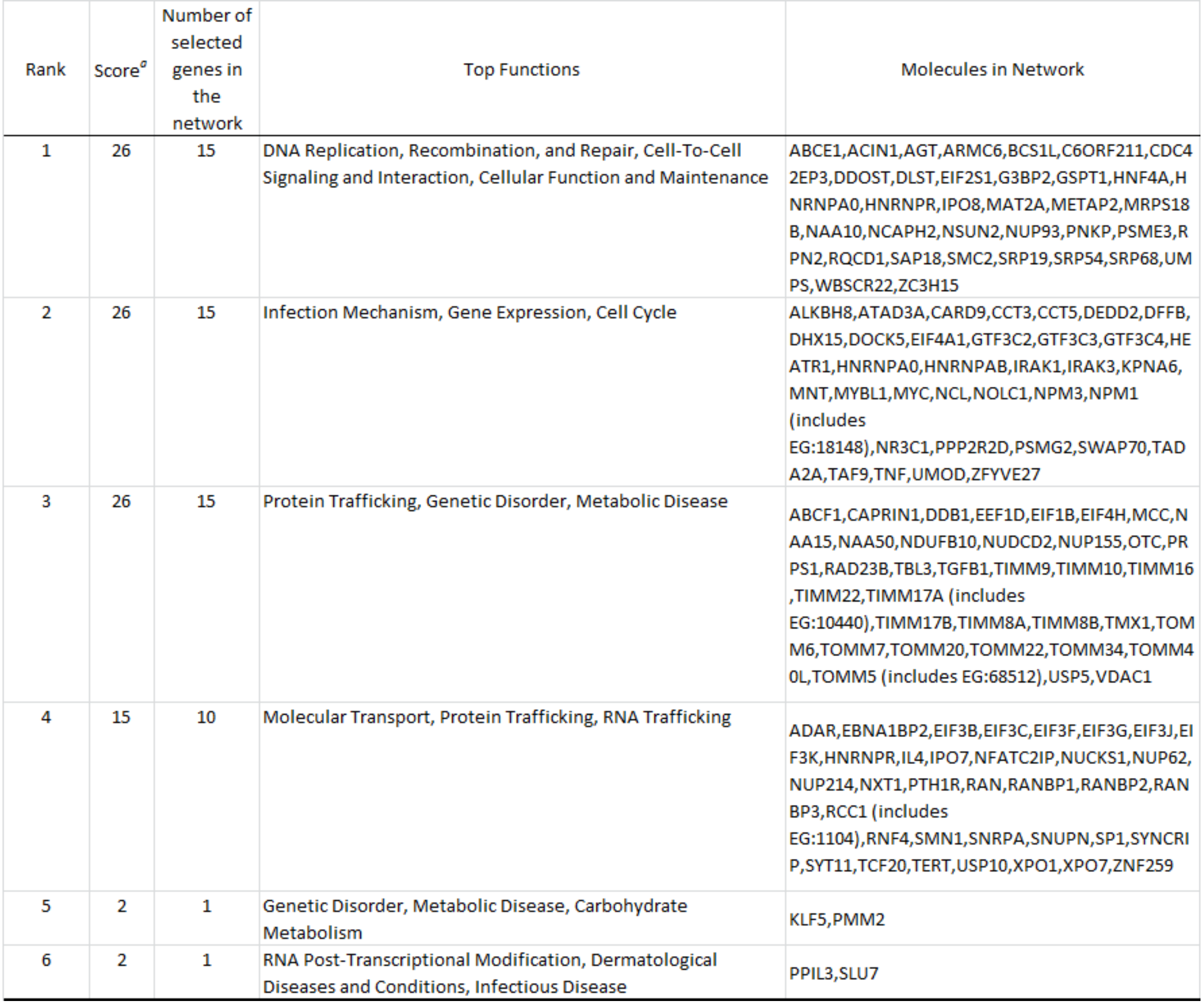
Table A1.2.IIIb. List of 40 biologically closest compound pairs and their hepatotoxicities. Nine pairs (23%) have opposite toxicities despite the similarity in gene expression values (0=dissimilar, 1=identical)[a]



No.	Compounds	Gene expression similarity	Observed hepatotoxici
71,82	nicotinic acid, enalapril maleate	0.94	0,0
48,126	ciprofloxacin, vancomycin	0.93	0,0
32,97	amiodarone, sulpiride	0.93	0,0
75,88	famotidine, triamterene	0.93	0,0
18115	diclofenac, ethanol	0.93	0,0
85,89	papaverine, disopyramide	0.93	0,1
72,77	erythromycin ethylsuccinate, chlorpheniramine	0.93	0,0
65,96	hydroxyzine, tolbutamide	0.93	0,0
47,60	lomustine, iproniazid	0.92	1,0
23,86	cyclophosphamide, penicillamine	0.92	1,0
102,107	triazolam, chlormadinone	0.92	0,0
39,76	labetalol, ranitidine	0.92	0,0
90,91	mexiletine, tiopronin	0.92	1,0
4,25	phenobarbital, phenytoin	0.92	0,0
5,41	valproic acid, glibenclamide	0.91	0,0
35,36	haloperidol, fluphenazine	0.91	0,0
56,57	monocrotaline, vitamin A	0.91	1,1
42,45	griseofulvin, ketoconazole	0.91	0,0
8,64	rifampicin, amitriptyline	0.91	0,0
84,92	caffeine, acetazolamide	0.91	0,0
19,52	nitrofurantoin, tamoxifen	0.91	0,0
67119	quinidine, cephalothin	0.91	0,0
22,28	diazepam, propylthiouracil	0.91	1,1
54,112	methyldopa, bromoethanamine	0.9	0,0
58,59	tacrine, moxisylyte	0.9	0,0
10,74	allyl alcohol, mefenamic acid	0.9	1,0
40,73	methyltestosterone, ethambutol	0.9	1,1
27,111	allopurinol, etoposide	0.9	0,0
49,68	pemoline, furosemide	0.89	0,0
55,61	methimazole, chloramphenicol	0.89	0,1
99,110	simvastatin, benziodarone	0.88	1,1
14,94	aspirin, promethazine	0.88	1,1
17,43	carbamazepine, flutamide	0.88	0,1
6,30	clofibrate, gemfibrozil	0.88	0,1
20,69	benzbromarone, fenofibrate	0.88	1,1
3,101	carbon tetrachloride, dantrolene	0.87	1,1
50,117	chlormezanone, bucetin	0.86	1,1
1,93	acetaminophen, disulfiram	0.86	1,1
24,118	methapyrilene, phenylanthranilic acid	0.85	1,0
114,124	nimesulide, ticlopidine	0.84	1,1

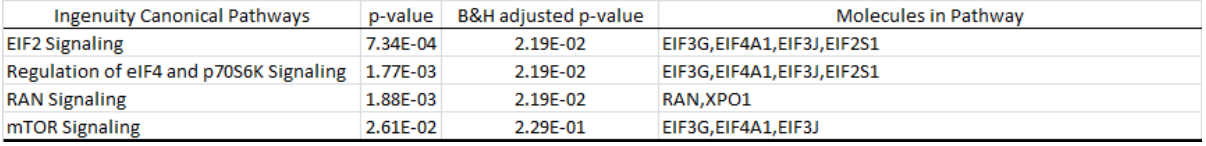
[a] Similarity = 1 - normalized pairwise Euclidean distance.
Identical pairs have a similarity of 1; completely different pairs have zero similarity.

Table A1.2.IVa. Molecular networks representing the predictors of hepatotoxicity (64 up-regulated genes)



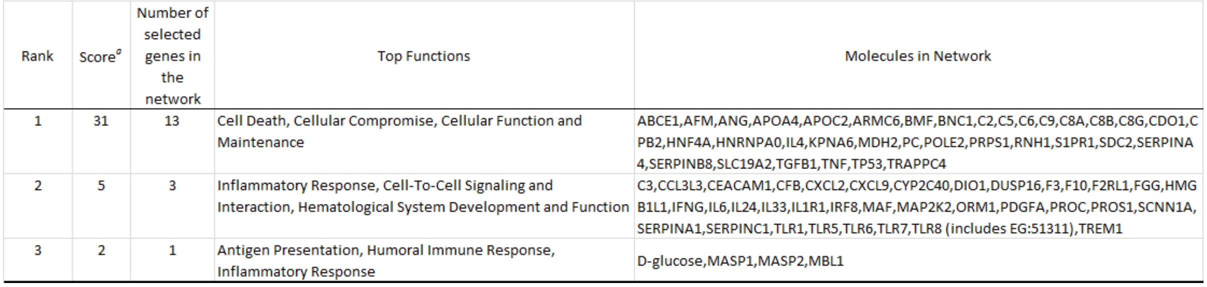
Rank	Score[a]	Number of selected genes in the network	Top Functions	Molecules in Network
1	26	15	DNA Replication, Recombination, and Repair, Cell-To-Cell Signaling and Interaction, Cellular Function and Maintenance	ABCE1,ACIN1,AGT,ARMC6,BCS1L,C6ORF211,CDC42EP3,DDOST,DLST,EIF2S1,G3BP2,GSPT1,HNF4A,HNRNPA0,HNRNPR,IPO8,MAT2A,METAP2,MRPS18B,NAA10,NCAPH2,NSUN2,NUP93,PNKP,PSME3,RPN2,RQCD1,SAP18,SMC2,SRP19,SRP54,SRP68,UMPS,WBSCR22,ZC3H15
2	26	15	Infection Mechanism, Gene Expression, Cell Cycle	ALKBH8,ATAD3A,CARD9,CCT3,CCT5,DEDD2,DFFB,DHX15,DOCK5,EIF4A1,GTF3C2,GTF3C3,GTF3C4,HEATR1,HNRNPA0,HNRNPAB,IRAK1,IRAK3,KPNA6,MNT,MYBL1,MYC,NCL,NOLC1,NPM3,NPM1 (includes EG:18148),NR3C1,PPP2R2D,PSMG2,SWAP70,TADA2A,TAF9,TNF,UMOD,ZFYVE27
3	26	15	Protein Trafficking, Genetic Disorder, Metabolic Disease	ABCF1,CAPRIN1,DDB1,EEF1D,EIF1B,EIF4H,MCC,NAA15,NAA50,NDUFB10,NUDCD2,NUP155,OTC,PRPS1,RAD23B,TBL3,TGFB1,TIMM9,TIMM10,TIMM16,TIMM22,TIMM17A (includes EG:10440),TIMM17B,TIMM8A,TIMM8B,TMX1,TOMM6,TOMM7,TOMM20,TOMM22,TOMM34,TOMM40L,TOMM5 (includes EG:68512),USP5,VDAC1
4	15	10	Molecular Transport, Protein Trafficking, RNA Trafficking	ADAR,EBNA1BP2,EIF3B,EIF3C,EIF3F,EIF3G,EIF3J,EIF3K,HNRNPR,IL4,IPO7,NFATC2IP,NUCKS1,NUP62,NUP214,NXT1,PTH1R,RAN,RANBP1,RANBP2,RANBP3,RCC1 (includes EG:1104),RNF4,SMN1,SNRPA,SNUPN,SP1,SYNCRIP,SYT11,TCF20,TERT,USP10,XPO1,XPO7,ZNF259
5	2	1	Genetic Disorder, Metabolic Disease, Carbohydrate Metabolism	KLF5,PMM2
6	2	1	RNA Post-Transcriptional Modification, Dermatological Diseases and Conditions, Infectious Disease	PPIL3,SLU7

Table A1.2.IVb. Canonical pathways representing the predictors of hepatotoxicity (64 up-regulated genes)



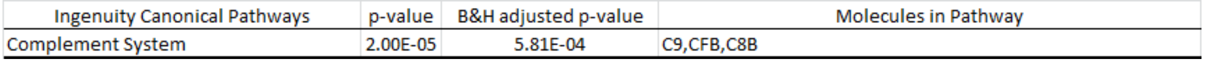
Ingenuity Canonical Pathways	p-value	B&H adjusted p-value	Molecules in Pathway
EIF2 Signaling	7.34E-04	2.19E-02	EIF3G,EIF4A1,EIF3J,EIF2S1
Regulation of eIF4 and p70S6K Signaling	1.77E-03	2.19E-02	EIF3G,EIF4A1,EIF3J,EIF2S1
RAN Signaling	1.88E-03	2.19E-02	RAN,XPO1
mTOR Signaling	2.61E-02	2.29E-01	EIF3G,EIF4A1,EIF3J

Table A1.2.IVc. Molecular networks representing the predictors of hepatotoxicity (21 down-regulated genes)



Rank	Score[a]	Number of selected genes in the network	Top Functions	Molecules in Network
1	31	13	Cell Death, Cellular Compromise, Cellular Function and Maintenance	ABCE1,AFM,ANG,APOA4,APOC2,ARMC6,BMF,BNC1,C2,C5,C6,C9,C8A,C8B,C8G,CDO1,CPB2,HNF4A,HNRNPA0,IL4,KPNA6,MDH2,PC,POLE2,PRPS1,RNH1,S1PR1,SDC2,SERPINA4,SERPINB8,SLC19A2,TGFB1,TNF,TP53,TRAPPC4
2	5	3	Inflammatory Response, Cell-To-Cell Signaling and Interaction, Hematological System Development and Function	C3,CCL3L3,CEACAM1,CFB,CXCL2,CXCL9,CYP2C40,DIO1,DUSP16,F3,F10,F2RL1,FGG,HMGB1L1,IFNG,IL6,IL24,IL33,IL1R1,IRF8,MAF,MAP2K2,ORM1,PDGFA,PROC,PROS1,SCNN1A,SERPINA1,SERPINC1,TLR1,TLR5,TLR6,TLR7,TLR8 (includes EG:51311),TREM1
3	2	1	Antigen Presentation, Humoral Immune Response, Inflammatory Response	D-glucose,MASP1,MASP2,MBL1

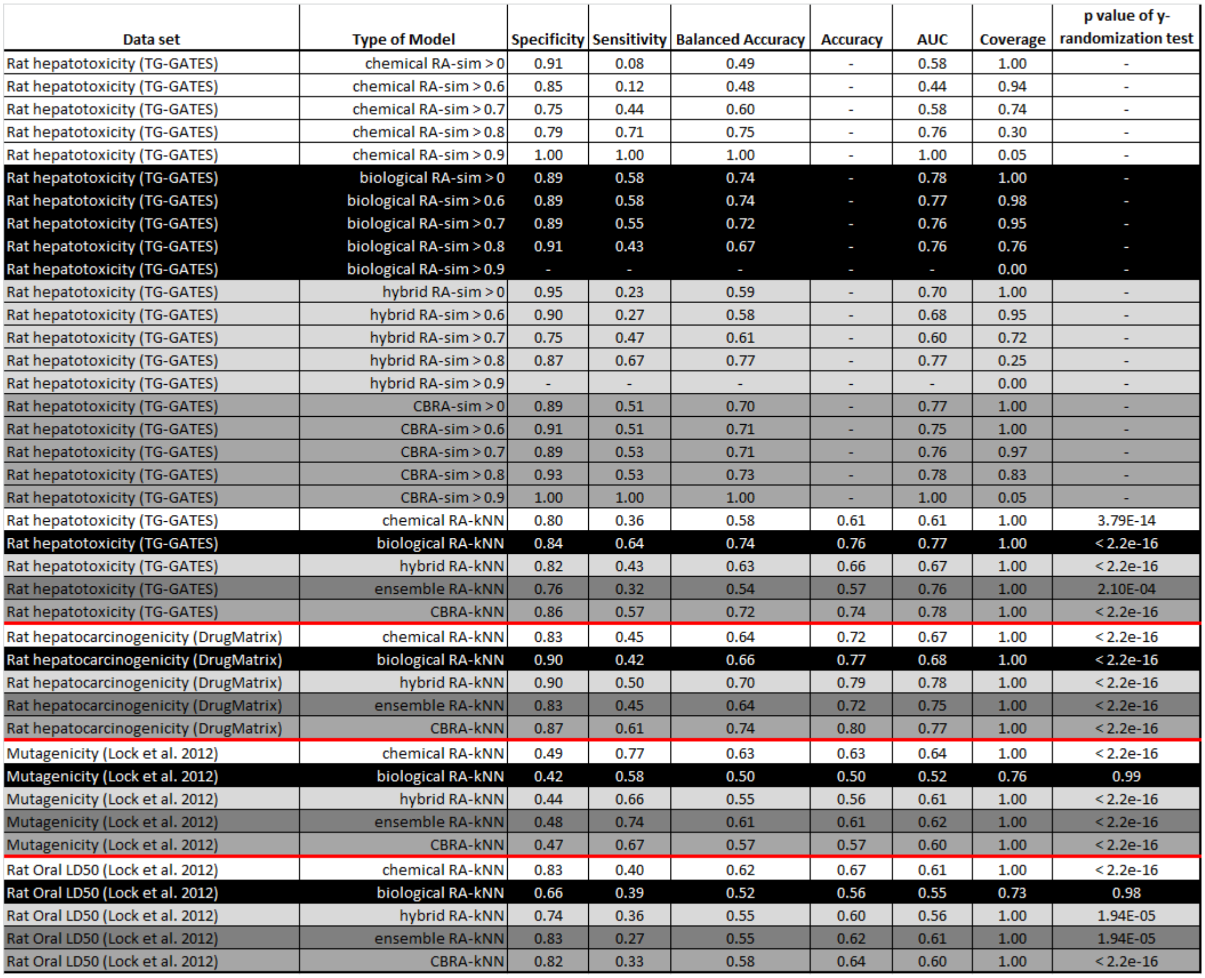
Table A1.2.IVd. Canonical pathways representing the predictors of hepatotoxicity (21 down-regulated genes)



Ingenuity Canonical Pathways	p-value	B&H adjusted p-value	Molecules in Pathway
Complement System	2.00E-05	5.81E-04	C9,CFB,C8B

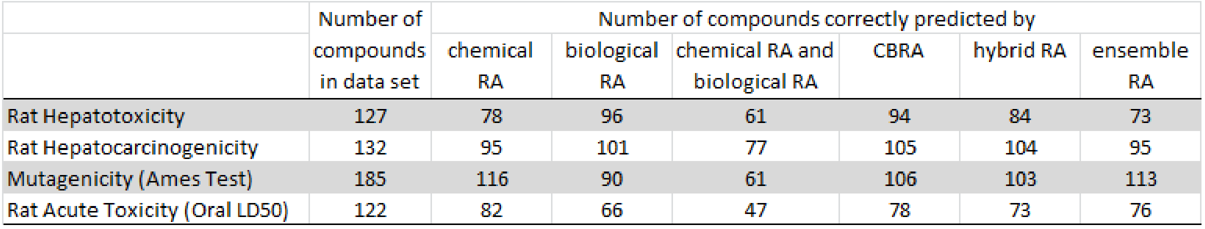
Supplemental tables for Chapter 3 (also available online at doi:10.1021/tx400110f)

Table A1.3.1. Model performance



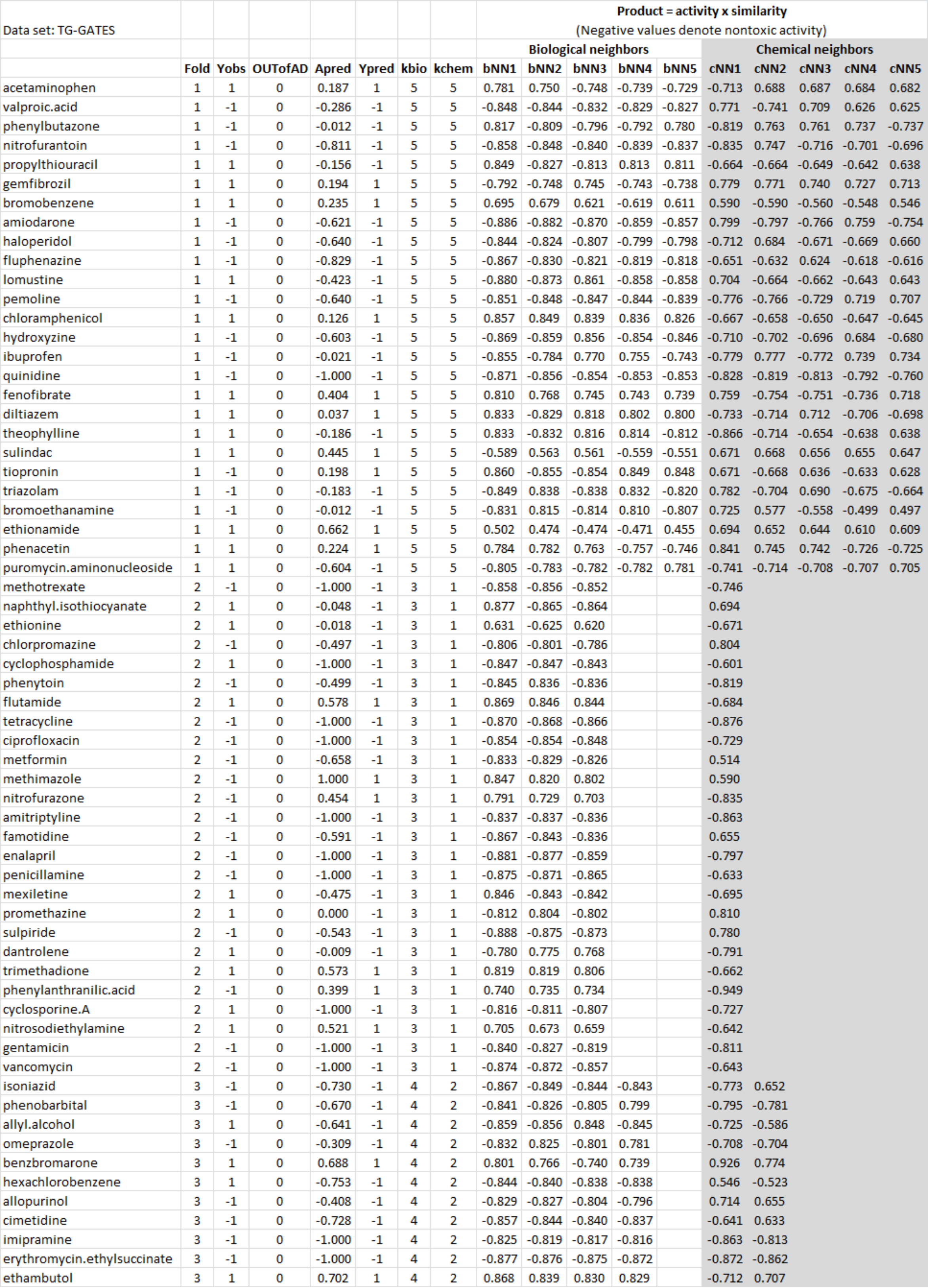
Data set	Type of Model	Specificity	Sensitivity	Balanced Accuracy	Accuracy	AUC	Coverage	p value of y-randomization test
Rat hepatotoxicity (TG-GATES)	chemical RA-sim > 0	0.91	0.08	0.49	-	0.58	1.00	-
Rat hepatotoxicity (TG-GATES)	chemical RA-sim > 0.6	0.85	0.12	0.48	-	0.44	0.94	-
Rat hepatotoxicity (TG-GATES)	chemical RA-sim > 0.7	0.75	0.44	0.60	-	0.58	0.74	-
Rat hepatotoxicity (TG-GATES)	chemical RA-sim > 0.8	0.79	0.71	0.75	-	0.76	0.30	-
Rat hepatotoxicity (TG-GATES)	chemical RA-sim > 0.9	1.00	1.00	1.00	-	1.00	0.05	-
Rat hepatotoxicity (TG-GATES)	biological RA-sim > 0	0.89	0.58	0.74	-	0.78	1.00	-
Rat hepatotoxicity (TG-GATES)	biological RA-sim > 0.6	0.89	0.58	0.74	-	0.77	0.98	-
Rat hepatotoxicity (TG-GATES)	biological RA-sim > 0.7	0.89	0.55	0.72	-	0.76	0.95	-
Rat hepatotoxicity (TG-GATES)	biological RA-sim > 0.8	0.91	0.43	0.67	-	0.76	0.76	-
Rat hepatotoxicity (TG-GATES)	biological RA-sim > 0.9	-	-	-	-	-	0.00	-
Rat hepatotoxicity (TG-GATES)	hybrid RA-sim > 0	0.95	0.23	0.59	-	0.70	1.00	-
Rat hepatotoxicity (TG-GATES)	hybrid RA-sim > 0.6	0.90	0.27	0.58	-	0.68	0.95	-
Rat hepatotoxicity (TG-GATES)	hybrid RA-sim > 0.7	0.75	0.47	0.61	-	0.60	0.72	-
Rat hepatotoxicity (TG-GATES)	hybrid RA-sim > 0.8	0.87	0.67	0.77	-	0.77	0.25	-
Rat hepatotoxicity (TG-GATES)	hybrid RA-sim > 0.9	-	-	-	-	-	0.00	-
Rat hepatotoxicity (TG-GATES)	CBRA-sim > 0	0.89	0.51	0.70	-	0.77	1.00	-
Rat hepatotoxicity (TG-GATES)	CBRA-sim > 0.6	0.91	0.51	0.71	-	0.75	1.00	-
Rat hepatotoxicity (TG-GATES)	CBRA-sim > 0.7	0.89	0.53	0.71	-	0.76	0.97	-
Rat hepatotoxicity (TG-GATES)	CBRA-sim > 0.8	0.93	0.53	0.73	-	0.78	0.83	-
Rat hepatotoxicity (TG-GATES)	CBRA-sim > 0.9	1.00	1.00	1.00	-	1.00	0.05	-
Rat hepatotoxicity (TG-GATES)	chemical RA-kNN	0.80	0.36	0.58	0.61	0.61	1.00	3.79E-14
Rat hepatotoxicity (TG-GATES)	biological RA-kNN	0.84	0.64	0.74	0.76	0.77	1.00	< 2.2e-16
Rat hepatotoxicity (TG-GATES)	hybrid RA-kNN	0.82	0.43	0.63	0.66	0.67	1.00	< 2.2e-16
Rat hepatotoxicity (TG-GATES)	ensemble RA-kNN	0.76	0.32	0.54	0.57	0.76	1.00	2.10E-04
Rat hepatotoxicity (TG-GATES)	CBRA-kNN	0.86	0.57	0.72	0.74	0.78	1.00	< 2.2e-16
Rat hepatocarcinogenicity (DrugMatrix)	chemical RA-kNN	0.83	0.45	0.64	0.72	0.67	1.00	< 2.2e-16
Rat hepatocarcinogenicity (DrugMatrix)	biological RA-kNN	0.90	0.42	0.66	0.77	0.68	1.00	< 2.2e-16
Rat hepatocarcinogenicity (DrugMatrix)	hybrid RA-kNN	0.90	0.50	0.70	0.79	0.78	1.00	< 2.2e-16
Rat hepatocarcinogenicity (DrugMatrix)	ensemble RA-kNN	0.83	0.45	0.64	0.72	0.75	1.00	< 2.2e-16
Rat hepatocarcinogenicity (DrugMatrix)	CBRA-kNN	0.87	0.61	0.74	0.80	0.77	1.00	< 2.2e-16
Mutagenicity (Lock et al. 2012)	chemical RA-kNN	0.49	0.77	0.63	0.63	0.64	1.00	< 2.2e-16
Mutagenicity (Lock et al. 2012)	biological RA-kNN	0.42	0.58	0.50	0.50	0.52	0.76	0.99
Mutagenicity (Lock et al. 2012)	hybrid RA-kNN	0.44	0.66	0.55	0.56	0.61	1.00	< 2.2e-16
Mutagenicity (Lock et al. 2012)	ensemble RA-kNN	0.48	0.74	0.61	0.61	0.62	1.00	< 2.2e-16
Mutagenicity (Lock et al. 2012)	CBRA-kNN	0.47	0.67	0.57	0.57	0.60	1.00	< 2.2e-16
Rat Oral LD50 (Lock et al. 2012)	chemical RA-kNN	0.83	0.40	0.62	0.67	0.61	1.00	< 2.2e-16
Rat Oral LD50 (Lock et al. 2012)	biological RA-kNN	0.66	0.39	0.52	0.56	0.55	0.73	0.98
Rat Oral LD50 (Lock et al. 2012)	hybrid RA-kNN	0.74	0.36	0.55	0.60	0.56	1.00	1.94E-05
Rat Oral LD50 (Lock et al. 2012)	ensemble RA-kNN	0.83	0.27	0.55	0.62	0.61	1.00	1.94E-05
Rat Oral LD50 (Lock et al. 2012)	CBRA-kNN	0.82	0.33	0.58	0.64	0.60	1.00	< 2.2e-16

Table A1.3.2. Number of compounds correctly predicted by various read-across (RA) models per data set



	Number of compounds in data set	Number of compounds correctly predicted by					
		chemical RA	biological RA	chemical RA and biological RA	CBRA	hybrid RA	ensemble RA
Rat Hepatotoxicity	127	78	96	61	94	84	73
Rat Hepatocarcinogenicity	132	95	101	77	105	104	95
Mutagenicity (Ames Test)	185	116	90	61	106	103	113
Rat Acute Toxicity (Oral LD50)	122	82	66	47	78	73	76

Table A1.3.3. Compounds and their chemical and biological neighbors in TG-GATES data set.



Data set: TG-GATES								Product = activity x similarity (Negative values denote nontoxic activity)									
								Biological neighbors					Chemical neighbors				
	Fold	Yobs	OUTofAD	Apred	Ypred	kbio	kchem	bNN1	bNN2	bNN3	bNN4	bNN5	cNN1	cNN2	cNN3	cNN4	cNN5
acetaminophen	1	1	0	0.187	1	5	5	0.781	0.750	-0.748	-0.739	-0.729	-0.713	0.688	0.687	0.684	0.682
valproic.acid	1	-1	0	-0.286	-1	5	5	-0.848	-0.844	-0.832	-0.829	-0.827	0.771	-0.741	0.709	0.626	0.625
phenylbutazone	1	-1	0	-0.012	-1	5	5	0.817	-0.809	-0.796	-0.792	0.780	-0.819	0.763	0.761	0.737	-0.737
nitrofurantoin	1	-1	0	-0.811	-1	5	5	-0.858	-0.848	-0.840	-0.839	-0.837	-0.835	0.747	-0.716	-0.701	-0.696
propylthiouracil	1	1	0	-0.156	-1	5	5	0.849	-0.827	-0.813	0.813	0.811	-0.664	-0.664	-0.649	-0.642	0.638
gemfibrozil	1	1	0	0.194	1	5	5	-0.792	-0.748	0.745	-0.743	-0.738	0.779	0.771	0.740	0.727	0.713
bromobenzene	1	1	0	0.235	1	5	5	0.695	0.679	0.621	-0.619	0.611	0.590	-0.590	-0.560	-0.548	0.546
amiodarone	1	-1	0	-0.621	-1	5	5	-0.886	-0.882	-0.870	-0.859	-0.857	0.799	-0.797	-0.766	0.759	-0.754
haloperidol	1	-1	0	-0.640	-1	5	5	-0.844	-0.824	-0.807	-0.799	-0.798	-0.712	0.684	-0.671	-0.669	0.660
fluphenazine	1	-1	0	-0.829	-1	5	5	-0.867	-0.830	-0.821	-0.819	-0.818	-0.651	-0.632	0.624	-0.618	-0.616
lomustine	1	1	0	-0.423	-1	5	5	-0.880	-0.873	0.861	-0.858	-0.858	0.704	-0.664	-0.662	-0.643	0.643
pemoline	1	-1	0	-0.640	-1	5	5	-0.851	-0.848	-0.847	-0.844	-0.839	-0.776	-0.766	-0.729	0.719	0.707
chloramphenicol	1	1	0	0.126	1	5	5	0.857	0.849	0.839	0.836	0.826	-0.667	-0.658	-0.650	-0.647	-0.645
hydroxyzine	1	-1	0	-0.603	-1	5	5	-0.869	-0.859	0.856	-0.854	-0.846	-0.710	-0.702	-0.696	0.684	-0.680
ibuprofen	1	-1	0	-0.021	-1	5	5	-0.855	-0.784	0.770	0.755	-0.743	-0.779	0.777	-0.772	0.739	0.734
quinidine	1	-1	0	-1.000	-1	5	5	-0.871	-0.856	-0.854	-0.853	-0.853	-0.828	-0.819	-0.813	-0.792	-0.760
fenofibrate	1	1	0	0.404	1	5	5	0.810	0.768	0.745	0.743	0.739	0.759	-0.754	-0.751	-0.736	0.718
diltiazem	1	1	0	0.037	1	5	5	0.833	-0.829	0.818	0.802	0.800	-0.733	-0.714	0.712	-0.706	-0.698
theophylline	1	1	0	-0.186	-1	5	5	0.833	-0.832	0.816	0.814	-0.812	-0.866	-0.714	-0.654	-0.638	0.638
sulindac	1	1	0	0.445	1	5	5	-0.589	0.563	0.561	-0.559	-0.551	0.671	0.668	0.656	0.655	0.647
tiopronin	1	-1	0	0.198	1	5	5	0.860	-0.855	-0.854	0.849	0.848	0.671	-0.668	0.636	-0.633	0.628
triazolam	1	-1	0	-0.183	-1	5	5	-0.849	0.838	-0.838	0.832	-0.820	0.782	-0.704	0.690	-0.675	-0.664
bromoethanamine	1	-1	0	-0.012	-1	5	5	-0.831	0.815	-0.814	0.810	-0.807	0.725	0.577	-0.558	-0.499	0.497
ethionamide	1	1	0	0.662	1	5	5	0.502	0.474	-0.474	-0.471	0.455	0.694	0.652	0.644	0.610	0.609
phenacetin	1	1	0	0.224	1	5	5	0.784	0.782	0.763	-0.757	-0.746	0.841	0.745	0.742	-0.726	-0.725
puromycin.aminonucleoside	1	1	0	-0.604	-1	5	5	-0.805	-0.783	-0.782	-0.782	0.781	-0.741	-0.714	-0.708	-0.707	0.705
methotrexate	2	-1	0	-1.000	-1	3	1	-0.858	-0.856	-0.852			-0.746				
naphthyl.isothiocyanate	2	1	0	-0.048	-1	3	1	0.877	-0.865	-0.864			0.694				
ethionine	2	1	0	-0.018	-1	3	1	0.631	-0.625	0.620			-0.671				
chlorpromazine	2	-1	0	-0.497	-1	3	1	-0.806	-0.801	-0.786			0.804				
cyclophosphamide	2	1	0	-1.000	-1	3	1	-0.847	-0.847	-0.843			-0.601				
phenytoin	2	-1	0	-0.499	-1	3	1	-0.845	0.836	-0.836			-0.819				
flutamide	2	1	0	0.578	1	3	1	0.869	0.846	0.844			-0.684				
tetracycline	2	-1	0	-1.000	-1	3	1	-0.870	-0.868	-0.866			-0.876				
ciprofloxacin	2	-1	0	-1.000	-1	3	1	-0.854	-0.854	-0.848			-0.729				
metformin	2	-1	0	-0.658	-1	3	1	-0.833	-0.829	-0.826			0.514				
methimazole	2	-1	0	1.000	1	3	1	0.847	0.820	0.802			0.590				
nitrofurazone	2	-1	0	0.454	1	3	1	0.791	0.729	0.703			-0.835				
amitriptyline	2	-1	0	-1.000	-1	3	1	-0.837	-0.837	-0.836			-0.863				
famotidine	2	-1	0	-0.591	-1	3	1	-0.867	-0.843	-0.836			0.655				
enalapril	2	-1	0	-1.000	-1	3	1	-0.881	-0.877	-0.859			-0.797				
penicillamine	2	-1	0	-1.000	-1	3	1	-0.875	-0.871	-0.865			-0.633				
mexiletine	2	1	0	-0.475	-1	3	1	0.846	-0.843	-0.842			-0.695				
promethazine	2	1	0	0.000	-1	3	1	-0.812	0.804	-0.802			0.810				
sulpiride	2	-1	0	-0.543	-1	3	1	-0.888	-0.875	-0.873			0.780				
dantrolene	2	1	0	-0.009	-1	3	1	-0.780	0.775	0.768			-0.791				
trimethadione	2	1	0	0.573	1	3	1	0.819	0.819	0.806			-0.662				
phenylanthranilic.acid	2	-1	0	0.399	1	3	1	0.740	0.735	0.734			-0.949				
cyclosporine.A	2	1	0	-1.000	-1	3	1	-0.816	-0.811	-0.807			-0.727				
nitrosodiethylamine	2	1	0	0.521	1	3	1	0.705	0.673	0.659			-0.642				
gentamicin	2	-1	0	-1.000	-1	3	1	-0.840	-0.827	-0.819			-0.811				
vancomycin	2	-1	0	-1.000	-1	3	1	-0.874	-0.872	-0.857			-0.643				
isoniazid	3	-1	0	-0.730	-1	4	2	-0.867	-0.849	-0.844	-0.843		-0.773	0.652			
phenobarbital	3	-1	0	-0.670	-1	4	2	-0.841	-0.826	-0.805	0.799		-0.795	-0.781			
allyl.alcohol	3	1	0	-0.641	-1	4	2	-0.859	-0.856	0.848	-0.845		-0.725	-0.586			
omeprazole	3	-1	0	-0.309	-1	4	2	-0.832	0.825	-0.801	0.781		-0.708	-0.704			
benzbromarone	3	1	0	0.688	1	4	2	0.801	0.766	-0.740	0.739		0.926	0.774			
hexachlorobenzene	3	1	0	-0.753	-1	4	2	-0.844	-0.840	-0.838	-0.838		0.546	-0.523			
allopurinol	3	-1	0	-0.408	-1	4	2	-0.829	-0.827	-0.804	-0.796		0.714	0.655			
cimetidine	3	-1	0	-0.728	-1	4	2	-0.857	-0.844	-0.840	-0.837		-0.641	0.633			
imipramine	3	-1	0	-1.000	-1	4	2	-0.825	-0.819	-0.817	-0.816		-0.863	-0.813			
erythromycin.ethylsuccinate	3	-1	0	-1.000	-1	4	2	-0.877	-0.876	-0.875	-0.872		-0.872	-0.862			
ethambutol	3	1	0	0.702	1	4	2	0.868	0.839	0.830	0.829		-0.712	0.707			

Table A1.3.3 (continued). Compounds and their chemical and biological neighbors in TG-GATES data set.

Data set: TG-GATES								Product = activity x similarity (Negative values denote nontoxic activity)									
								Biological neighbors					Chemical neighbors				
	Fold	Yobs	OUTofAD	Apred	Ypred	kbio	kchem	bNN1	bNN2	bNN3	bNN4	bNN5	cNN1	cNN2	cNN3	cNN4	cNN5
chlorpheniramine	3	-1	0	-1.000	-1	4	2	-0.862	-0.858	-0.848	-0.848		-0.704	-0.703			
tannic.acid	3	-1	0	-0.700	-1	4	2	-0.835	-0.832	-0.824	-0.823		0.710	-0.705			
captopril	3	-1	0	-0.726	-1	4	2	-0.867	-0.857	-0.854	-0.851		-0.668	0.649			
caffeine	3	-1	0	-0.641	-1	4	2	-0.836	-0.830	-0.814	-0.812		0.866	-0.663			
disopyramide	3	1	0	-1.000	-1	4	2	-0.874	-0.872	-0.860	-0.855		-0.763	-0.753			
acetazolamide	3	-1	0	-0.705	-1	4	2	-0.825	-0.816	-0.807	-0.803		0.664	-0.583			
disulfiram	3	1	0	0.297	1	4	2	0.760	-0.759	-0.735	0.734		0.638	0.628			
colchicine	3	1	0	-0.121	-1	4	2	0.570	0.569	0.552	-0.552		-0.835	-0.769			
tolbutamide	3	-1	0	-0.327	-1	4	2	-0.854	-0.854	0.837	-0.834		0.806	-0.699			
clomipramine	3	1	0	-1.000	-1	4	2	-0.853	-0.842	-0.835	-0.832		-0.804	-0.782			
danazol	3	-1	0	0.677	1	4	2	0.820	0.747	-0.745	0.730		0.808	0.761			
bendazac	3	1	0	-0.335	-1	4	2	-0.793	0.781	0.765	-0.748		-0.791	-0.774			
etoposide	3	-1	0	-0.674	-1	4	2	-0.818	-0.815	-0.803	-0.799		0.776	-0.758			
nimesulide	3	1	0	0.684	1	4	2	0.764	0.741	0.733	0.725		-0.683	0.681			
clofibrate	4	-1	0	0.449	1	5	2	0.787	0.777	-0.742	0.740	0.736	0.736	-0.695			
diazepam	4	1	0	-0.147	-1	5	2	-0.809	0.805	0.794	-0.793	-0.788	0.704	-0.704			
methapyrilene	4	1	0	0.464	1	5	2	0.774	0.771	0.757	-0.743	0.741	0.810	-0.667			
coumarin	4	1	0	0.428	1	5	2	-0.696	0.677	0.676	0.675	0.672	0.804	-0.709			
sulfasalazine	4	1	0	-0.705	-1	5	2	-0.813	0.808	-0.802	-0.790	-0.788	-0.780	-0.696			
glibenclamide	4	-1	0	-0.755	-1	5	2	-0.846	-0.830	-0.817	-0.814	-0.813	-0.696	0.673			
ethinylestradiol	4	1	0	0.439	1	5	2	0.813	0.791	0.782	0.782	-0.780	0.911	-0.808			
monocrotaline	4	1	0	-0.144	-1	5	2	0.809	-0.803	0.789	-0.786	-0.785	-0.705	0.705			
tacrine	4	-1	0	-0.410	-1	5	2	0.790	-0.784	0.781	-0.779	-0.775	-0.712	-0.703			
moxisylyte	4	-1	0	-0.702	-1	5	2	0.809	-0.796	-0.780	-0.775	-0.769	-0.758	-0.746			
furosemide	4	-1	0	-0.470	-1	5	2	-0.818	-0.816	-0.808	0.802	-0.802	-0.741	0.635			
mefenamic.acid	4	-1	0	-0.168	-1	5	2	0.787	0.782	-0.781	-0.769	0.769	-0.949	-0.781			
nifedipine	4	-1	0	0.442	1	5	2	0.801	0.751	0.744	0.737	0.730	-0.739	-0.719			
papaverine	4	-1	0	-0.140	-1	5	2	-0.853	0.844	0.839	-0.830	-0.826	0.835	-0.828			
triamterene	4	-1	0	-0.448	-1	5	2	0.833	-0.827	-0.815	-0.810	-0.807	-0.716	0.681			
acarbose	4	-1	0	-0.722	-1	5	2	-0.805	-0.790	-0.786	0.780	-0.776	-0.872	-0.809			
simvastatin	4	1	0	-0.139	-1	5	2	-0.771	0.761	0.750	-0.743	-0.743	0.784	-0.776			
ajmaline	4	-1	0	0.140	1	5	2	-0.823	-0.820	0.806	0.804	0.800	0.713	-0.711			
lornoxicam	4	-1	0	0.116	1	5	2	-0.744	0.743	-0.742	-0.741	0.738	0.670	0.662			
chlormadinone	4	-1	0	0.142	1	5	2	-0.800	0.792	-0.785	0.783	0.782	0.751	-0.750			
benziodarone	4	1	0	0.158	1	5	2	-0.778	0.766	0.754	0.748	-0.744	0.926	-0.799			
cephalothin	4	-1	0	-0.436	-1	5	2	-0.802	-0.794	0.792	-0.791	-0.790	0.733	-0.704			
acetamidofluorene	4	1	0	0.684	1	5	2	0.702	0.656	0.655	0.651	0.631	0.767	-0.762			
ticlopidine	4	1	0	0.153	1	5	2	0.806	0.792	-0.763	0.753	-0.731	-0.733	0.679			
doxorubicin	4	-1	0	-1.000	-1	5	2	-0.837	-0.820	-0.810	-0.806	-0.802	-0.876	-0.862			
carbon.tetrachloride	5	1	0	0.006	1	5	4	-0.781	0.772	0.768	-0.768	0.766	0.394	-0.380	-0.373	-0.365	
rifampicin	5	-1	0	-0.769	-1	5	4	0.864	-0.861	-0.837	-0.833	-0.827	-0.832	-0.818	-0.816	-0.798	
aspirin	5	1	0	0.115	1	5	4	-0.815	0.813	-0.808	0.803	0.795	0.804	-0.771	-0.714	0.704	
thioacetamide	5	1	0	-0.061	1	5	4	0.706	-0.693	0.684	-0.681	0.679	-0.577	0.539	-0.513	-0.480	
carbamazepine	5	-1	0	-0.099	1	5	4	0.790	0.783	0.776	0.770	-0.767	-0.813	-0.766	-0.737	-0.721	
diclofenac	5	-1	0	-0.807	-1	5	4	-0.873	-0.864	-0.862	-0.861	-0.860	-0.730	-0.712	0.690	-0.679	
WY-14643	5	1	0	0.148	1	5	4	0.798	0.776	-0.761	0.757	0.741	-0.666	-0.663	-0.662	0.638	
thioridazine	5	-1	0	-0.606	-1	5	4	-0.827	-0.822	-0.821	-0.821	-0.819	-0.701	0.699	-0.651	0.640	
adapin	5	-1	0	-0.796	-1	5	4	-0.871	-0.868	-0.850	-0.850	-0.849	-0.813	-0.792	-0.776	0.756	
labetalol	5	-1	0	-0.786	-1	5	4	-0.869	-0.840	-0.838	-0.834	-0.834	-0.813	-0.797	0.791	-0.766	
methyltestosterone	5	1	0	0.571	1	5	4	0.824	0.806	0.802	-0.791	0.791	0.911	0.784	0.764	-0.761	
griseofulvin	5	-1	0	-0.583	-1	5	4	-0.829	-0.829	-0.826	-0.824	-0.817	0.754	-0.750	0.720	-0.719	
perhexiline	5	-1	0	-0.801	-1	5	4	-0.818	-0.810	-0.799	-0.799	-0.794	-0.766	-0.743	-0.703	0.688	
ketoconazole	5	-1	0	-0.373	-1	5	4	-0.858	-0.839	0.837	-0.829	-0.827	-0.696	0.658	-0.650	0.648	
chlormezanone	5	1	0	0.610	1	5	4	0.789	0.748	0.732	0.731	0.728	0.661	0.622	-0.611	-0.602	
tamoxifen	5	-1	0	-0.321	-1	5	4	-0.865	0.846	0.834	-0.828	0.828	-0.819	-0.797	-0.792	-0.774	
methyldopa	5	-1	0	-0.553	-1	5	4	-0.822	-0.821	0.816	-0.806	-0.804	0.713	-0.695	-0.687	-0.676	
vitamin.A	5	1	0	0.335	1	5	4	0.847	0.846	0.828	-0.826	-0.826	0.791	0.784	0.779	-0.777	
iproniazid	5	-1	0	-0.182	-1	5	4	-0.865	-0.855	-0.854	0.854	-0.853	-0.773	0.688	0.686	0.678	
chlorpropamide	5	1	0	-0.651	-1	5	4	-0.861	-0.852	-0.852	-0.851	-0.841	-0.806	-0.673	0.609	0.605	
nicotinic.acid	5	-1	0	-0.132	-1	5	4	-0.893	-0.876	-0.872	0.872	0.866	-0.741	0.683	0.662	-0.637	
ranitidine	5	-1	0	-0.609	-1	5	4	-0.878	-0.875	-0.869	-0.863	-0.862	0.706	-0.701	-0.699	0.691	
terbinafine	5	1	0	0.365	1	5	4	0.823	0.821	0.816	0.815	0.810	-0.771	-0.755	-0.695	0.691	
ethanol	5	-1	0	-0.528	-1	5	4	0.879	-0.860	-0.858	-0.853	-0.849	0.586	-0.558	-0.381	-0.378	
bucetin	5	1	0	0.348	1	5	4	0.767	0.756	-0.742	0.741	-0.740	0.841	0.758	0.750	-0.748	

Table A1.3.4. Genes and their local importance scores

Probe ID (rat2302 chip)	Global rank[a]	Global importance[b]	Local importance (chloramphenicol)	Local importance (carbamazepine)	Local importance (benzbromarone)	Gene symbol	Gene name	Gene ontology (GO) terms (as determined by GO Consortium; retrieved by BioConductor)
1390326_at	1	0.03415	0.00787	0.00615	0.00140	NA	NA	NA
1389470_at	2	0.03294	0.00154	0.00000	0.00000	*Cfb*	complement factor B	complement activation; complement activation, alternative pathway; response to lipopolysaccharide; extracellular space
1369206_at	3	0.02896	0.00267	0.00293	0.00521	*Cpb2*	carboxypeptidase B2 (plasma)	metabolic process; blood coagulation; proteolysis; response to drug; response to heat; fibrinolysis; negative regulation of fibrinolysis; extracellular region; extracellular space; carboxypeptidase activity
1387765_at	4	0.02076	0.00896	0.00483	-0.00077	*Mbl1*	mannose-binding lectin (protein A) 1	complement activation, lectin pathway; complement activation, classical pathway; defense response to Gram-positive bacterium; negative regulation of growth of symbiont in host; positive regulation of phagocytosis; killing by host of symbiont cells; extracellular region; collagen; extracellular space; calcium ion binding
1368461_at	5	0.01636	0.00000	0.00302	-0.00014	*Slc22a8*	solute carrier family 22 (organic anion transporter), member 8	response to toxin; quaternary ammonium group transport; organic anion transport; response to methotrexate; glutathione transport; plasma membrane; integral to membrane; basolateral plasma membrane; protein kinase C binding; transmembrane transporter activity
1368741_at	6	0.01608	0.00198	0.00000	0.00000	*C9*	complement component 9	blood coagulation; induction of apoptosis; activation of caspase activity; immune response; complement activation, alternative pathway; complement activation, classical pathway; cytolysis; plasma membrane; integral to membrane; extracellular region
1388582_at	7	0.01576	0.00000	0.00282	0.00000	*Psme3*	proteasome (prosome, macropain) activator subunit 3	regulation of proteasomal protein catabolic process; positive regulation of endopeptidase activity; regulation of apoptosis; p53 binding; endopeptidase activator activity; MDM2 binding
1382343_at	8	0.01562	0.00410	0.00000	0.00000	NA	NA	NA
1398832_at	9	0.01559	0.00177	0.00263	0.00666	*Ncl*	nucleolin	angiogenesis; ribonucleoprotein complex; nucleus; cytoplasm; nucleoplasm; nucleolus; cell cortex; nucleotide binding; nucleic acid binding; DNA binding
1377676_at	10	0.01485	0.00000	-0.00184	0.00000	*Nucks1*	nuclear casein kinase and cyclin-dependent kinase substrate 1	regulation of cell cycle; nucleus; cytoplasm; double-stranded DNA binding; single-stranded DNA binding
1394991_at	11	0.01282	0.00227	0.00000	0.00000	*Irak1*	interleukin-1 receptor-associated kinase 1	regulation of cytokine-mediated signaling pathway; signal transduction; protein phosphorylation; activation of NF-kappaB-inducing kinase activity; JNK cascade; negative regulation of transcription, DNA-dependent; cytokine-mediated signaling pathway; lipopolysaccharide-mediated signaling pathway; negative regulation of NF-kappaB transcription factor activity; response to peptidoglycan
1371876_at	12	0.01049	0.00553	0.00000	0.00000	*Psmg2*	proteasome (prosome, macropain) assembly chaperone 2	regulation of cell cycle; apoptosis; mitotic cell cycle spindle assembly checkpoint; regulation of apoptosis; proteasome assembly; nucleus
1371266_at	13	0.00938	0.00000	-0.00372	-0.00265	*Afm*	afamin	vitamin transport; extracellular region; extracellular space; vitamin E binding
1388810_at	14	0.00904	0.00750	0.00987	0.00571	*Abce1*	ATP-binding cassette, subfamily E (OABP), member 1	mitochondrion; cytoplasm; electron carrier activity; nucleotide binding; ATP binding; ATPase activity; iron-sulfur cluster binding
1385756_at	15	0.00899	0.00000	-0.00101	0.00000	*Itih1*	inter-alpha trypsin inhibitor, heavy chain 1	NA
1398259_at	16	0.00788	0.00000	0.00562	0.00000	*Nup155*	nucleoporin 155	nucleocytoplasmic transport; protein transport; mRNA transport; transmembrane transport; membrane; nucleus; nuclear pore; nuclear membrane; structural constituent of nuclear pore
1395173_at	17	0.00778	-0.00004	-0.00265	0.00000	*Caprin1*	cell cycle associated protein 1	negative regulation of translation; cell differentiation; positive regulation of dendrite morphogenesis; positive regulation of dendritic spine morphogenesis; cytoplasmic mRNA processing body; cytosol; cytoplasm; stress granule; dendrite; cell projection
1372150_at	18	0.00747	0.00000	0.00199	0.00000	*Usp10*	ubiquitin specific peptidase 10	DNA repair; ubiquitin-dependent protein catabolic process; protein deubiquitination; DNA damage response, signal transduction by p53 class mediator; nucleus; cytoplasm; early endosome; intermediate filament cytoskeleton; p53 binding; cysteine-type peptidase activity
1388974_at	19	0.00513	0.00333	0.00000	0.00000	*Srp19*	signal recognition particle 19	SRP-dependent cotranslational protein targeting to membrane; response to drug; mitochondrion; nucleus; cytoplasm; nucleolus; signal recognition particle, endoplasmic reticulum targeting; 7S RNA binding
1389587_at	20	0.00466	0.00489	0.00000	0.00000	*Umps*	uridine monophosphate synthetase	'de novo' pyrimidine base biosynthetic process; pyrimidine nucleotide biosynthetic process; UMP biosynthetic process; female pregnancy; lactation; nucleoside metabolic process; cellular response to drug; 'de novo' UMP biosynthetic process; soluble fraction; nucleus
1388425_at	21	0.00466	0.00000	-0.00103	-0.00110	*Oaf*	OAF homolog (Drosophila)	NA
1371980_at	22	0.00326	-0.00088	0.00000	0.00000	*Atad3a*	ATPase family, AAA domain containing 3A	mitochondrion; mitochondrial inner membrane; nucleotide binding; nucleoside-triphosphatase activity; ATP binding
1372067_at	23	0.00318	0.00000	0.00000	0.00000	*Tmx1*	thioredoxin-related transmembrane protein 1	cell redox homeostasis; anti-apoptosis; endoplasmic reticulum; membrane fraction; endoplasmic reticulum membrane; disulfide oxidoreductase activity
1367590_at	24	0.00313	0.00000	0.00000	0.00997	*Ran*	RAN, member RAS oncogene family	microtubule cytoskeleton organization; protein import into nucleus; protein export from nucleus; nucleocytoplasmic transport; spermatid development; chromatin; nucleus; cytoplasm; protein complex; GTPase activity
1392544_at	25	0.00283	0.00000	-0.00037	0.00000	*Rqcd1*	rcd1 (required for cell differentiation) homolog 1 (S. pombe)	regulation of transcription, DNA-dependent; cytokine-mediated signaling pathway; cytoplasmic mRNA processing body; nucleus; cytoplasm; binding
1367713_at	26	0.00280	0.00284	0.00000	0.00094	*Eif2s1*	eukaryotic translation initiation factor 2, subunit 1 alpha	translation; translational initiation; regulation of translation; regulation of translational initiation in response to stress; protein autophosphorylation; cytosol; nucleus; cytoplasm; eukaryotic translation initiation factor 2 complex; eukaryotic translation initiation factor 2B complex
1389450_at	27	0.00202	0.00000	0.00000	0.00000	NA	NA	NA
1372939_at	28	0.00152	0.00000	0.00396	0.00000	*Nudcd2*	NudC domain containing 2	intracellular
1398876_at	29	0.00101	0.00000	0.00643	0.00000	*Abcf1*	ATP-binding cassette, subfamily F (GCN20), member 1	ribosome biogenesis; translation; translational initiation; positive regulation of translation; ribosome; nucleus; nuclear envelope; cytoplasm; nucleoplasm; polysomal ribosome
1367755_at	30	-0.00036	0.00499	0.01326	-0.00020	*Cdo1*	cysteine dioxygenase, type I	cysteine metabolic process; lactation; response to organic nitrogen; L-cysteine catabolic process; response to glucagon stimulus; response to amino acid stimulus; response to ethanol; L-cysteine metabolic process; response to glucocorticoid stimulus; response to cAMP
1389110_at	31	-0.00061	0.00000	0.00000	0.00419	*LOC100360017*	mitochondrial ribosomal protein S6-like	translation; mitochondrion; ribosome; structural constituent of ribosome; rRNA binding
1367473_at	32	-0.00067	0.00277	0.00000	0.00088	*Tomm22*	translocase of outer mitochondrial membrane 22 homolog (yeast)	intracellular protein transport; protein import into mitochondrial outer membrane; transmembrane transport; integral to membrane of membrane fraction; membrane; mitochondrion; integral to membrane; mitochondrial outer membrane translocase complex; mitochondrial inner membrane; protein binding
1374886_at	33	-0.00162	0.00000	0.00000	0.00986	*Bcs1l*	BCS1-like (yeast)	mitochondrion organization; biological_process; mitochondrial respiratory chain complex I assembly; mitochondrial respiratory chain complex IV assembly; mitochondrial respiratory chain complex III assembly; mitochondrion; cellular_component; nucleotide binding; molecular_function; nucleoside-triphosphatase activity
1372519_at	34	-0.00182	0.00000	0.00000	0.00348	*Nup93*	nucleoporin 93	transport; protein transport; mRNA transport; transmembrane transport; nucleus; nuclear pore
1388134_at	35	-0.00212	0.00000	0.00000	0.00000	*Eef1d*	eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein)	signal transduction; positive regulation of I-kappaB kinase/NF-kappaB cascade; cytosol; eukaryotic translation elongation factor 1 complex; translation elongation factor activity; signal transducer activity
1386917_at	36	-0.00258	0.00000	0.00000	0.00000	*Pc*	pyruvate carboxylase	pyruvate metabolic process; gluconeogenesis; oxaloacetate metabolic process; lipid biosynthetic process; mitochondrion; soluble fraction; mitochondrial inner membrane; mitochondrial matrix; nucleotide binding; biotin carboxylase activity
1371939_at	37	-0.00280	0.00416	0.00000	0.00000	*Caprin1*	cell cycle associated protein 1	negative regulation of translation; cell differentiation; positive regulation of dendrite morphogenesis; positive regulation of dendritic spine morphogenesis; cytoplasmic mRNA processing body; cytosol; cytoplasm; stress granule; dendrite; cell projection
1388381_at	38	-0.00292	0.00000	0.00000	0.00000	*Eif3g*	eukaryotic translation initiation factor 3, subunit G	translational initiation; cytosol; nucleus; cytoplasm; eukaryotic translation initiation factor 3 complex; perinuclear region of cytoplasm; nucleotide binding; nucleic acid binding; RNA binding; translation initiation factor activity
1392465_at	39	-0.00326	0.00382	0.00000	0.00000	*Sap18*	Sin3-associated polypeptide 18	biological_process; cellular_component; molecular_function
1391491_a_at	40	-0.00394	0.00000	0.00000	0.00000	*Rad23b*	RAD23 homolog B (S. cerevisiae)	nucleotide-excision repair, DNA damage recognition; nucleotide-excision repair, DNA damage removal; nucleotide-excision repair; response to DNA damage stimulus; spermatogenesis; regulation of proteasomal ubiquitin-dependent protein catabolic process; proteasome complex; nucleus; cytoplasm; nucleoplasm
1371626_at	41	-0.00396	0.00467	0.00000	0.00000	*Srp68*	signal recognition particle 68	response to drug; cytoplasm; signal recognition particle, endoplasmic reticulum targeting; signal recognition particle binding
1372046_at	42	-0.00424	0.00000	0.00000	0.00000	*Gtf3c3*	general transcription factor IIIC, polypeptide 3	transcription factor TFIIIC complex; binding

[a] Global rank sorts the descriptors in order of decreasing global importance.

[b] Global importance measures the decrease in overall external balanced accuracy when the gene is permuted. Unlike the local importance score I(x,compound) which depends on the target compound, global importance G(x) measures gene x's contribution to the overall balanced accuracy of the model.

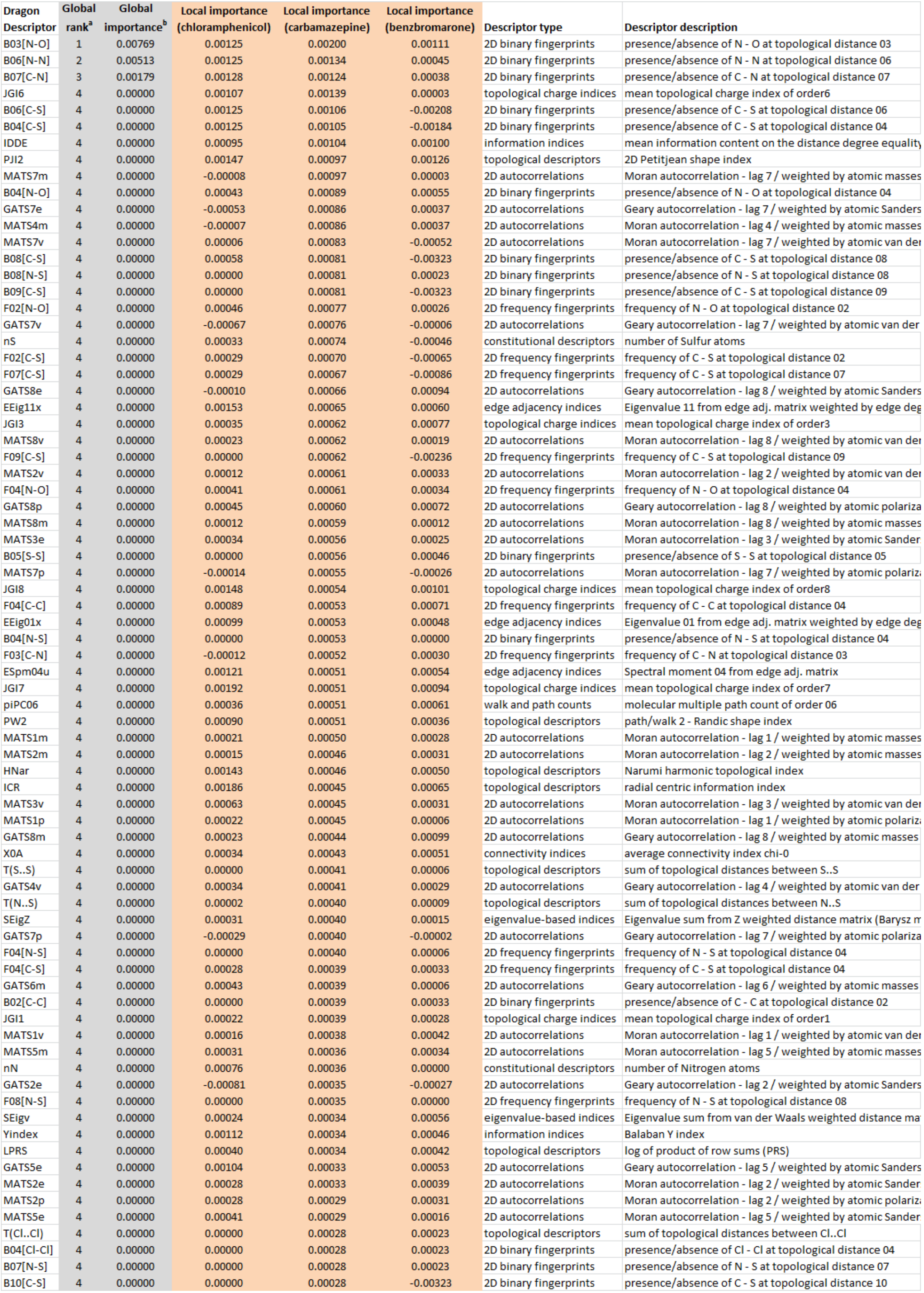
Table A1.3.4 (continued). Genes and their local importance scores

Probe ID (rat2302 chip)	Global rank[a]	Global importance[b]	Local importance (chloramphenicol)	Local importance (carbamazepine)	Local importance (benzbromarone)	Gene symbol	Gene name	Gene ontology (GO) terms (as determined by GO Consortium; retrieved by BioConductor)
1368165_at	43	-0.00466	0.00037	0.00000	0.00000	*Prps1*	phosphoribosyl pyrophosphate synthetase 1	5-phosphoribose 1-diphosphate biosynthetic process; purine base metabolic process; purine nucleotide biosynthetic process; AMP biosynthetic process; nervous system development; nucleoside metabolic process; ribonucleoside monophosphate biosynthetic process; nucleotide biosynthetic process; ribose phosphate metabolic process; organ regeneration
1370836_at	44	-0.00470	0.00204	0.00000	0.00142	*Serpina4*	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 4	negative regulation of endopeptidase activity; regulation of proteolysis; extracellular region; extracellular space; serine-type endopeptidase inhibitor activity
1388938_at	45	-0.00553	0.00000	0.00000	0.00000	*Usp5*	ubiquitin specific peptidase 5 (isopeptidase T)	ubiquitin-dependent protein catabolic process; positive regulation of proteasomal ubiquitin-dependent protein catabolic process; protein K48-linked deubiquitination; cysteine-type peptidase activity; ubiquitin thiolesterase activity; peptidase activity; metal ion binding; omega peptidase activity; zinc ion binding
1383514_s_at	46	-0.00606	0.00000	0.00000	0.00000	*Narg1*	NMDA receptor regulated 1	metabolic process; N-terminal protein amino acid acetylation; positive regulation of transcription, DNA-dependent; nucleus; cytoplasm; transcription factor complex; binding; N-acetyltransferase activity; acetyltransferase activity; ribosome binding
1372661_at	47	-0.00616	0.00000	0.00000	0.00696	*Tbl3*	transducin (beta)-like 3	rRNA processing; nucleus; nucleolus; small-subunit processome
1398757_at	48	-0.00653	0.00000	0.00457	0.00338	*Npm1*	nucleophosmin (nucleolar phosphoprotein B23, numatrin)	ribosomal large subunit export from nucleus; ribosomal small subunit export from nucleus; cell growth; regulation of cell cycle; nucleosome assembly; protein localization; DNA repair; rRNA export from nucleus; cell volume homeostasis; nucleocytoplasmic transport
1388107_at	49	-0.00677	0.00000	0.00000	0.00891	*Ppp2r2d*	protein phosphatase 2, regulatory subunit B, delta isoform	signal transduction; cell cycle; mitosis; exit from mitosis; cell division; protein phosphatase type 2A complex; cytosol; cytoplasm; nucleoplasm; phosphatidate phosphatase activity
1369902_at	50	-0.00697	0.00245	0.01165	0.00000	*Bmf*	Bcl2 modifying factor	apoptosis; biological_process; regulation of apoptosis; positive regulation of protein homooligomerization; positive regulation of release of cytochrome c from mitochondria; plasma membrane; acrosomal vesicle; cytosol; cytoplasm; actin cytoskeleton
1398937_at	51	-0.00717	0.00000	0.00338	0.00000	*Dhx15*	DEAH (Asp-Glu-Ala-His) box polypeptide 15	U12-type spliceosomal complex; nucleotide binding; nucleic acid binding; helicase activity; ATP binding; ATP-dependent helicase activity; hydrolase activity
1371109_at	52	-0.00727	0.00084	0.00091	-0.00012	*C8b*	complement component 8, beta polypeptide	immune response; complement activation, alternative pathway; complement activation, classical pathway; cytolysis; extracellular region; membrane attack complex; extracellular space; protein complex binding
1370785_s_at	53	-0.00737	0.00000	0.00000	0.00948	*Tomm20*	translocase of outer mitochondrial membrane 20 homolog (yeast)	protein targeting to mitochondrion; membrane; mitochondrion; integral to membrane; mitochondrial envelope; mitochondrial outer membrane translocase complex; unfolded protein binding; P-P-bond-hydrolysis-driven protein transmembrane transporter activity
1382923_at	54	-0.00737	0.00000	0.00000	0.00000	*Syncrip*	synaptotagmin binding, cytoplasmic RNA interacting protein	mRNA processing; RNA splicing; CRD-mediated mRNA stabilization; ribonucleoprotein complex; nucleus; cytoplasm; nucleoplasm; spliceosomal complex; CRD-mediated mRNA stability complex; catalytic step 2 spliceosome
1370166_at	55	-0.00788	0.00000	0.00000	0.00000	*Sdc2*	syndecan 2	response to hypoxia; response to caffeine; wound healing; dendrite morphogenesis; membrane; integral to membrane; neuronal cell body; synapse; protein binding; cytoskeletal protein binding
1370495_s_at	56	-0.00839	0.00157	0.00221	-0.00348	*Cyp2c13*	cytochrome P450, family 2, subfamily c, polypeptide 13	membrane; endoplasmic reticulum; endoplasmic reticulum membrane; microsome; electron carrier activity; monooxygenase activity; metal ion binding; oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; heme binding; aromatase activity
1373380_at	57	-0.00889	0.00000	0.00000	0.00000	*Zc3h15*	zinc finger CCCH-type containing 15	cytokine-mediated signaling pathway; plasma membrane; nucleus; cytoplasm; nucleolus; nucleic acid binding; metal ion binding; zinc ion binding
1379376_at	58	-0.00891	0.00000	0.00251	0.00000	NA	NA	NA
1375686_at	59	-0.00917	0.00000	0.00000	0.00000	*Ppil3*	peptidylprolyl isomerase (cyclophilin)-like 3	protein folding; mRNA processing; RNA splicing; spliceosomal complex; catalytic step 2 spliceosome; peptidyl-prolyl cis-trans isomerase activity; isomerase activity
1369852_at	60	-0.00923	0.00457	0.00118	0.00123	*F10*	coagulation factor X	blood coagulation; proteolysis; blood coagulation, extrinsic pathway; positive regulation of protein kinase B signaling cascade; plasma membrane; extracellular region; membrane fraction; microsome; calcium ion binding; serine-type endopeptidase activity
1368400_at	61	-0.00943	0.00000	0.00000	0.00000	*Timm8a1*	translocase of inner mitochondrial membrane 8 homolog a1 (yeast)	protein targeting to mitochondrion; protein transport; protein import into mitochondrial inner membrane; transmembrane transport; membrane; mitochondrion;
1371840_at	62	-0.00951	0.00000	0.00000	0.00000	*S1pr1*	sphingosine-1-phosphate receptor 1	angiogenesis; G-protein coupled receptor protein signaling pathway; inhibition of adenylate cyclase activity by G-protein signaling pathway; brain development; positive regulation of cell proliferation; transmission of nerve impulse; regulation of cell adhesion; neuron differentiation; positive regulation of cell migration; positive regulation of Ras GTPase activity
1389021_at	63	-0.00970	0.00000	0.00000	0.00000	*Asnsd1*	asparagine synthetase domain containing 1	asparagine biosynthetic process; asparagine synthase (glutamine-hydrolyzing) activity
1388397_at	64	-0.00985	0.00000	0.00000	0.00000	*Ebna1bp2*	EBNA1 binding protein 2	nucleus; nucleolus
1376570_at	65	-0.01063	0.00000	0.00000	0.00000	*Cct5*	chaperonin containing Tcp1, subunit 5 (epsilon)	protein folding; response to virus; cytosol; cytoplasm; nucleolus; microtubule organizing center; chaperonin-containing T-complex; cytoskeleton; unfolded protein binding; nucleotide binding
1374943_at	66	-0.01115	0.01042	0.00000	0.00000	*RGD1311378*	similar to RIKEN cDNA 2010011I20	NA
1390863_at	67	-0.01128	0.00000	0.00223	-0.00133	*Slc19a2*	solute carrier family 19 (thiamine transporter), member 2	biological_process; thiamine transport; plasma membrane; cellular_component; molecular_function; thiamine transmembrane transporter activity; thiamine uptake transmembrane transporter activity
1393139_at	68	-0.01222	0.00000	0.00000	-0.00022	*Apoc2*	apolipoprotein C-II	response to drug; positive regulation of phospholipase activity; positive regulation of triglyceride catabolic process; negative regulation of very-low-density lipoprotein particle clearance; negative regulation of cholesterol transport; cholesterol efflux; phospholipid efflux; chylomicron remnant clearance; high-density lipoprotein particle clearance; lipoprotein transport
1378551_at	69	-0.01226	0.00000	0.00129	0.00349	*Cyp20a1*	cytochrome P450, family 20, subfamily a, polypeptide 1	oxidation-reduction process; membrane; integral to membrane; electron carrier activity; monooxygenase activity; metal ion binding; oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; heme binding
1368362_a_at	70	-0.01244	0.00245	0.00000	0.00000	*Asgr2*	asialoglycoprotein receptor 2	endocytosis; glycoprotein metabolic process; bone mineralization; regulation of protein stability; lipid homeostasis; plasma membrane; membrane; integral to membrane; nucleolus; focal adhesion
1388150_at	71	-0.01279	0.00000	0.00616	0.00000	*Xpo1*	exportin 1, CRM1 homolog (yeast)	negative regulation of transcription from RNA polymerase II promoter; protein export from nucleus; intracellular protein transport; response to drug; regulation of centrosome duplication; protein localization to nucleus; regulation of protein catabolic process; regulation of protein export from nucleus; mRNA transport; kinetochore
1370304_at	72	-0.01431	0.00000	0.00000	0.00000	*Timm17a*	translocase of inner mitochondrial membrane 17 homolog A (yeast)	intracellular protein transport; transmembrane transport; membrane; mitochondrion; integral to membrane; mitochondrial inner membrane; mitochondrial inner
1370309_a_at	73	-0.01455	0.00000	0.00000	0.00056	*Hnrnpab*	heterogeneous nuclear ribonucleoprotein A/B	epithelial to mesenchymal transition; positive regulation of transcription, DNA-dependent; negative regulation of transcription, DNA-dependent; ribonucleoprotein complex; nucleus; cytoplasm; DNA binding; DNA replication origin binding; sequence-specific DNA binding transcription factor activity; sequence-specific DNA binding
1388424_at	74	-0.01545	0.00000	0.00000	0.00000	*Eif3j*	eukaryotic translation initiation factor 3, subunit J	cytosol; cytoplasm; eukaryotic translation initiation factor 3 complex; translation initiation factor activity
1371403_at	75	-0.01549	0.00000	0.00000	0.00353	*Cct3*	chaperonin containing Tcp1, subunit 3 (gamma)	protein folding; plasma membrane; cytoplasm; chaperonin-containing T-complex; unfolded protein binding; nucleotide binding; ATP binding
1385804_x_at	76	-0.01610	0.00304	0.00251	0.00000	*LOC688637*	similar to WD repeat domain 36	rRNA processing; small-subunit processome
1391280_at	77	-0.01869	0.00000	0.00000	-0.00058	*Mcts2*	malignant T cell amplified sequence 2	NA
1371936_at	78	-0.01900	0.00380	0.01998	0.01113	*Eif4a1*	eukaryotic translation initiation factor 4A, isoform 1	organ regeneration; cellular_component; molecular_function
1383685_at	79	-0.02131	0.00000	0.00000	0.00000	*Heatr1*	HEAT repeat containing 1	binding
1397458_at	80	-0.02222	0.00000	0.00000	0.00000	*Pmm2*	phosphomannomutase 2	dolichol-linked oligosaccharide biosynthetic process; biological_process; mannose biosynthetic process; cellular_component; cytoplasm; neuronal cell body; molecular_function; catalytic activity; phosphomannomutase activity

[a] Global rank sorts the descriptors in order of decreasing global importance.

[b] Global importance measures the decrease in overall external balanced accuracy when the gene is permuted. Unlike the local importance score I(x,compound) which depends on the target compound, global importance G(x) measures gene x's contribution to the overall balanced accuracy of the model.

Table A1.3.5. Chemical descriptors and their local importance scores



Dragon Descriptor	Global rank[a]	Global importance[b]	Local importance (chloramphenicol)	Local importance (carbamazepine)	Local importance (benzbromarone)	Descriptor type	Descriptor description
B03[N-O]	1	0.00769	0.00125	0.00200	0.00111	2D binary fingerprints	presence/absence of N - O at topological distance 03
B06[N-N]	2	0.00513	0.00125	0.00134	0.00045	2D binary fingerprints	presence/absence of N - N at topological distance 06
B07[C-N]	3	0.00179	0.00128	0.00124	0.00038	2D binary fingerprints	presence/absence of C - N at topological distance 07
JGI6	4	0.00000	0.00107	0.00139	0.00003	topological charge indices	mean topological charge index of order6
B06[C-S]	4	0.00000	0.00125	0.00106	-0.00208	2D binary fingerprints	presence/absence of C - S at topological distance 06
B04[C-S]	4	0.00000	0.00125	0.00105	-0.00184	2D binary fingerprints	presence/absence of C - S at topological distance 04
IDDE	4	0.00000	0.00095	0.00104	0.00100	information indices	mean information content on the distance degree equality
PJI2	4	0.00000	0.00147	0.00097	0.00126	topological descriptors	2D Petitjean shape index
MATS7m	4	0.00000	-0.00008	0.00097	0.00003	2D autocorrelations	Moran autocorrelation - lag 7 / weighted by atomic masses
B04[N-O]	4	0.00000	0.00043	0.00089	0.00055	2D binary fingerprints	presence/absence of N - O at topological distance 04
GATS7e	4	0.00000	-0.00053	0.00086	0.00037	2D autocorrelations	Geary autocorrelation - lag 7 / weighted by atomic Sanders
MATS4m	4	0.00000	-0.00007	0.00086	0.00037	2D autocorrelations	Moran autocorrelation - lag 4 / weighted by atomic masses
MATS7v	4	0.00000	0.00006	0.00083	-0.00052	2D autocorrelations	Moran autocorrelation - lag 7 / weighted by atomic van der
B08[C-S]	4	0.00000	0.00058	0.00081	-0.00323	2D binary fingerprints	presence/absence of C - S at topological distance 08
B08[N-S]	4	0.00000	0.00000	0.00081	0.00023	2D binary fingerprints	presence/absence of N - S at topological distance 08
B09[C-S]	4	0.00000	0.00000	0.00081	-0.00323	2D binary fingerprints	presence/absence of C - S at topological distance 09
F02[N-O]	4	0.00000	0.00046	0.00077	0.00026	2D frequency fingerprints	frequency of N - O at topological distance 02
GATS7v	4	0.00000	-0.00067	0.00076	-0.00006	2D autocorrelations	Geary autocorrelation - lag 7 / weighted by atomic van der
nS	4	0.00000	0.00033	0.00074	-0.00046	constitutional descriptors	number of Sulfur atoms
F02[C-S]	4	0.00000	0.00029	0.00070	-0.00065	2D frequency fingerprints	frequency of C - S at topological distance 02
F07[C-S]	4	0.00000	0.00029	0.00067	-0.00086	2D frequency fingerprints	frequency of C - S at topological distance 07
GATS8e	4	0.00000	-0.00010	0.00066	0.00094	2D autocorrelations	Geary autocorrelation - lag 8 / weighted by atomic Sanders
EEig11x	4	0.00000	0.00153	0.00065	0.00060	edge adjacency indices	Eigenvalue 11 from edge adj. matrix weighted by edge deg
JGI3	4	0.00000	0.00035	0.00062	0.00077	topological charge indices	mean topological charge index of order3
MATS8v	4	0.00000	0.00023	0.00062	0.00019	2D autocorrelations	Moran autocorrelation - lag 8 / weighted by atomic van der
F09[C-S]	4	0.00000	0.00000	0.00062	-0.00236	2D frequency fingerprints	frequency of C - S at topological distance 09
MATS2v	4	0.00000	0.00012	0.00061	0.00033	2D autocorrelations	Moran autocorrelation - lag 2 / weighted by atomic van der
F04[N-O]	4	0.00000	0.00041	0.00061	0.00034	2D frequency fingerprints	frequency of N - O at topological distance 04
GATS8p	4	0.00000	0.00045	0.00060	0.00072	2D autocorrelations	Geary autocorrelation - lag 8 / weighted by atomic polariza
MATS8m	4	0.00000	0.00012	0.00059	0.00012	2D autocorrelations	Moran autocorrelation - lag 8 / weighted by atomic masses
MATS3e	4	0.00000	0.00034	0.00056	0.00025	2D autocorrelations	Moran autocorrelation - lag 3 / weighted by atomic Sander
B05[S-S]	4	0.00000	0.00000	0.00056	0.00046	2D binary fingerprints	presence/absence of S - S at topological distance 05
MATS7p	4	0.00000	-0.00014	0.00055	-0.00026	2D autocorrelations	Moran autocorrelation - lag 7 / weighted by atomic polariz
JGI8	4	0.00000	0.00148	0.00054	0.00101	topological charge indices	mean topological charge index of order8
F04[C-C]	4	0.00000	0.00089	0.00053	0.00071	2D frequency fingerprints	frequency of C - C at topological distance 04
EEig01x	4	0.00000	0.00099	0.00053	0.00048	edge adjacency indices	Eigenvalue 01 from edge adj. matrix weighted by edge deg
B04[N-S]	4	0.00000	0.00000	0.00053	0.00000	2D binary fingerprints	presence/absence of N - S at topological distance 04
F03[C-N]	4	0.00000	-0.00012	0.00052	0.00030	2D frequency fingerprints	frequency of C - N at topological distance 03
ESpm04u	4	0.00000	0.00121	0.00051	0.00054	edge adjacency indices	Spectral moment 04 from edge adj. matrix
JGI7	4	0.00000	0.00192	0.00051	0.00094	topological charge indices	mean topological charge index of order7
piPC06	4	0.00000	0.00036	0.00051	0.00061	walk and path counts	molecular multiple path count of order 06
PW2	4	0.00000	0.00090	0.00051	0.00036	topological descriptors	path/walk 2 - Randic shape index
MATS1m	4	0.00000	0.00021	0.00050	0.00028	2D autocorrelations	Moran autocorrelation - lag 1 / weighted by atomic masses
MATS2m	4	0.00000	0.00015	0.00046	0.00031	2D autocorrelations	Moran autocorrelation - lag 2 / weighted by atomic masses
HNar	4	0.00000	0.00143	0.00046	0.00050	topological descriptors	Narumi harmonic topological index
ICR	4	0.00000	0.00186	0.00045	0.00065	topological descriptors	radial centric information index
MATS3v	4	0.00000	0.00063	0.00045	0.00031	2D autocorrelations	Moran autocorrelation - lag 3 / weighted by atomic van der
MATS1p	4	0.00000	0.00022	0.00045	0.00006	2D autocorrelations	Moran autocorrelation - lag 1 / weighted by atomic polariz
GATS8m	4	0.00000	0.00023	0.00044	0.00099	2D autocorrelations	Geary autocorrelation - lag 8 / weighted by atomic masses
X0A	4	0.00000	0.00034	0.00043	0.00051	connectivity indices	average connectivity index chi-0
T(S..S)	4	0.00000	0.00000	0.00041	0.00006	topological descriptors	sum of topological distances between S..S
GATS4v	4	0.00000	0.00034	0.00041	0.00029	2D autocorrelations	Geary autocorrelation - lag 4 / weighted by atomic van der
T(N..S)	4	0.00000	0.00002	0.00040	0.00009	topological descriptors	sum of topological distances between N..S
SEigZ	4	0.00000	0.00031	0.00040	0.00015	eigenvalue-based indices	Eigenvalue sum from Z weighted distance matrix (Barysz m
GATS7p	4	0.00000	-0.00029	0.00040	-0.00002	2D autocorrelations	Geary autocorrelation - lag 7 / weighted by atomic polariza
F04[N-S]	4	0.00000	0.00000	0.00040	0.00006	2D frequency fingerprints	frequency of N - S at topological distance 04
F04[C-S]	4	0.00000	0.00028	0.00039	0.00033	2D frequency fingerprints	frequency of C - S at topological distance 04
GATS6m	4	0.00000	0.00043	0.00039	0.00006	2D autocorrelations	Geary autocorrelation - lag 6 / weighted by atomic masses
B02[C-C]	4	0.00000	0.00000	0.00039	0.00033	2D binary fingerprints	presence/absence of C - C at topological distance 02
JGI1	4	0.00000	0.00022	0.00039	0.00028	topological charge indices	mean topological charge index of order1
MATS1v	4	0.00000	0.00016	0.00038	0.00042	2D autocorrelations	Moran autocorrelation - lag 1 / weighted by atomic van der
MATS5m	4	0.00000	0.00031	0.00036	0.00034	2D autocorrelations	Moran autocorrelation - lag 5 / weighted by atomic masses
nN	4	0.00000	0.00076	0.00036	0.00000	constitutional descriptors	number of Nitrogen atoms
GATS2e	4	0.00000	-0.00081	0.00035	-0.00027	2D autocorrelations	Geary autocorrelation - lag 2 / weighted by atomic Sanders
F08[N-S]	4	0.00000	0.00000	0.00035	0.00000	2D frequency fingerprints	frequency of N - S at topological distance 08
SEigv	4	0.00000	0.00024	0.00034	0.00056	eigenvalue-based indices	Eigenvalue sum from van der Waals weighted distance ma
Yindex	4	0.00000	0.00112	0.00034	0.00046	information indices	Balaban Y index
LPRS	4	0.00000	0.00040	0.00034	0.00042	topological descriptors	log of product of row sums (PRS)
GATS5e	4	0.00000	0.00104	0.00033	0.00053	2D autocorrelations	Geary autocorrelation - lag 5 / weighted by atomic Sanders
MATS2e	4	0.00000	0.00028	0.00033	0.00039	2D autocorrelations	Moran autocorrelation - lag 2 / weighted by atomic Sander
MATS2p	4	0.00000	0.00028	0.00029	0.00031	2D autocorrelations	Moran autocorrelation - lag 2 / weighted by atomic polariz
MATS5e	4	0.00000	0.00041	0.00029	0.00016	2D autocorrelations	Moran autocorrelation - lag 5 / weighted by atomic Sander
T(Cl..Cl)	4	0.00000	0.00000	0.00028	0.00023	topological descriptors	sum of topological distances between Cl..Cl
B04[Cl-Cl]	4	0.00000	0.00000	0.00028	0.00023	2D binary fingerprints	presence/absence of Cl - Cl at topological distance 04
B07[N-S]	4	0.00000	0.00000	0.00028	0.00023	2D binary fingerprints	presence/absence of N - S at topological distance 07
B10[C-S]	4	0.00000	0.00000	0.00028	-0.00323	2D binary fingerprints	presence/absence of C - S at topological distance 10

[a] Global rank sorts the descriptors in order of decreasing global importance.

[b] Global importance measures the decrease in overall external balanced accuracy when the descriptor is permuted. Unlike the local importance score I(x,compound) which depends on the target compound, global importance G(x) measures descriptor x's contribution to the overall balanced accuracy of the model.

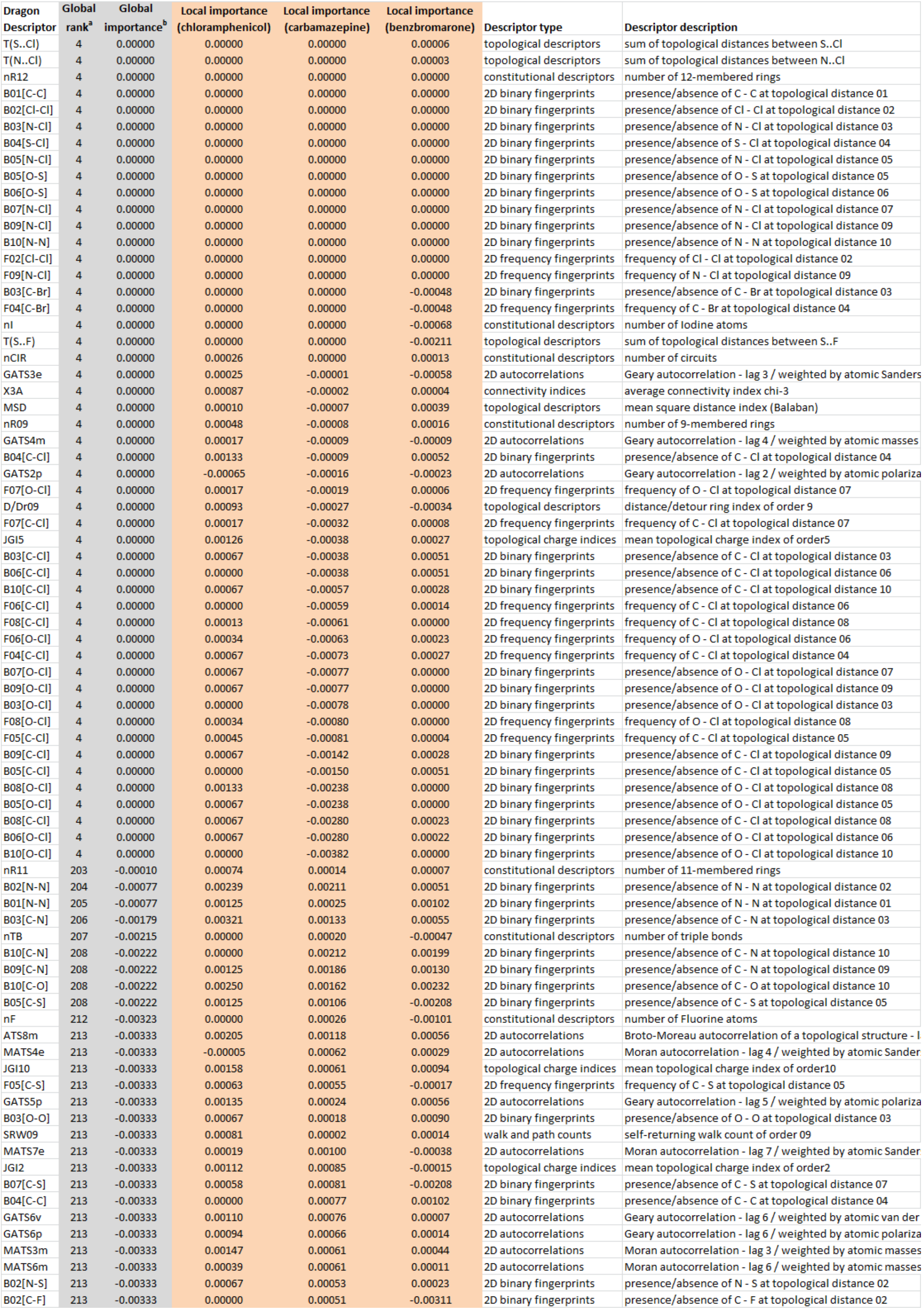
Table A1.3.5 (continued). Chemical descriptors and their local importance scores

Dragon Descriptor	Global rank[a]	Global importance[b]	Local importance (chloramphenicol)	Local importance (carbamazepine)	Local importance (benzbromarone)	Descriptor type	Descriptor description
nR08	4	0.00000	0.00000	0.00028	0.00000	constitutional descriptors	number of 8-membered rings
B02[S-S]	4	0.00000	0.00000	0.00028	0.00023	2D binary fingerprints	presence/absence of S - S at topological distance 02
B03[S-S]	4	0.00000	0.00000	0.00028	0.00023	2D binary fingerprints	presence/absence of S - S at topological distance 03
B02[C-Br]	4	0.00000	0.00000	0.00028	-0.00025	2D binary fingerprints	presence/absence of C - Br at topological distance 02
F07[O-O]	4	0.00000	0.00031	0.00027	0.00055	2D frequency fingerprints	frequency of O - O at topological distance 07
B03[N-N]	4	0.00000	0.00250	0.00027	0.00124	2D binary fingerprints	presence/absence of N - N at topological distance 03
MATS5v	4	0.00000	0.00042	0.00027	0.00015	2D autocorrelations	Moran autocorrelation - lag 5 / weighted by atomic van der
F05[N-O]	4	0.00000	0.00005	0.00026	0.00018	2D frequency fingerprints	frequency of N - O at topological distance 05
VEA1	4	0.00000	0.00073	0.00026	0.00043	eigenvalue-based indices	eigenvector coefficient sum from adjacency matrix
X1A	4	0.00000	0.00098	0.00026	0.00011	connectivity indices	average connectivity index chi-1
B01[N-S]	4	0.00000	0.00067	0.00026	0.00023	2D binary fingerprints	presence/absence of N - S at topological distance 01
B06[N-S]	4	0.00000	0.00058	0.00026	0.00113	2D binary fingerprints	presence/absence of N - S at topological distance 06
B01[O-S]	4	0.00000	0.00058	0.00026	-0.00323	2D binary fingerprints	presence/absence of O - S at topological distance 01
B09[N-F]	4	0.00000	0.00000	0.00026	0.00022	2D binary fingerprints	presence/absence of N - F at topological distance 09
B07[N-N]	4	0.00000	0.00000	0.00026	0.00000	2D binary fingerprints	presence/absence of N - N at topological distance 07
B09[N-N]	4	0.00000	0.00000	0.00026	0.00000	2D binary fingerprints	presence/absence of N - N at topological distance 09
B10[N-S]	4	0.00000	0.00000	0.00026	0.00000	2D binary fingerprints	presence/absence of N - S at topological distance 10
B10[O-S]	4	0.00000	0.00000	0.00026	0.00000	2D binary fingerprints	presence/absence of O - S at topological distance 10
B10[C-F]	4	0.00000	0.00000	0.00026	-0.00324	2D binary fingerprints	presence/absence of C - F at topological distance 10
B07[O-F]	4	0.00000	0.00000	0.00026	-0.00340	2D binary fingerprints	presence/absence of O - F at topological distance 07
F02[C-N]	4	0.00000	0.00002	0.00024	0.00017	2D frequency fingerprints	frequency of C - N at topological distance 02
J	4	0.00000	0.00024	0.00023	0.00060	topological descriptors	Balaban distance connectivity index
GATS7m	4	0.00000	-0.00027	0.00023	0.00018	2D autocorrelations	Geary autocorrelation - lag 7 / weighted by atomic masses
MATS6e	4	0.00000	0.00053	0.00023	-0.00003	2D autocorrelations	Moran autocorrelation - lag 6 / weighted by atomic Sander
Xindex	4	0.00000	0.00021	0.00023	0.00052	information indices	Balaban X index
nCL	4	0.00000	0.00023	0.00022	0.00030	constitutional descriptors	number of Chlorine atoms
F04[N-N]	4	0.00000	0.00062	0.00022	0.00000	2D frequency fingerprints	frequency of N - N at topological distance 04
PCD	4	0.00000	0.00041	0.00021	0.00022	walk and path counts	difference between multiple path count and path count
F03[O-O]	4	0.00000	0.00021	0.00021	0.00026	2D frequency fingerprints	frequency of O - O at topological distance 03
X4A	4	0.00000	0.00061	0.00020	0.00023	connectivity indices	average connectivity index chi-4
GATS3v	4	0.00000	-0.00018	0.00019	-0.00016	2D autocorrelations	Geary autocorrelation - lag 3 / weighted by atomic van der
F04[C-N]	4	0.00000	0.00030	0.00019	0.00010	2D frequency fingerprints	frequency of C - N at topological distance 04
GATS1e	4	0.00000	0.00044	0.00019	0.00019	2D autocorrelations	Geary autocorrelation - lag 1 / weighted by atomic Sanders
F02[C-Cl]	4	0.00000	0.00011	0.00018	0.00034	2D frequency fingerprints	frequency of C - Cl at topological distance 02
F06[O-O]	4	0.00000	0.00005	0.00018	0.00031	2D frequency fingerprints	frequency of O - O at topological distance 06
GATS1p	4	0.00000	0.00036	0.00018	-0.00036	2D autocorrelations	Geary autocorrelation - lag 1 / weighted by atomic polariza
F03[N-O]	4	0.00000	0.00005	0.00017	0.00026	2D frequency fingerprints	frequency of N - O at topological distance 03
GATS1v	4	0.00000	0.00029	0.00016	0.00007	2D autocorrelations	Geary autocorrelation - lag 1 / weighted by atomic van der
F06[C-N]	4	0.00000	0.00043	0.00015	0.00028	2D frequency fingerprints	frequency of C - N at topological distance 06
F02[O-O]	4	0.00000	0.00038	0.00015	0.00008	2D frequency fingerprints	frequency of O - O at topological distance 02
GATS2m	4	0.00000	-0.00052	0.00015	0.00016	2D autocorrelations	Geary autocorrelation - lag 2 / weighted by atomic masses
Uindex	4	0.00000	0.00021	0.00015	0.00009	information indices	Balaban U index
PW3	4	0.00000	0.00089	0.00013	-0.00004	topological descriptors	path/walk 3 - Randic shape index
F10[N-O]	4	0.00000	0.00000	0.00013	0.00029	2D frequency fingerprints	frequency of N - O at topological distance 10
PW4	4	0.00000	0.00148	0.00012	0.00046	topological descriptors	path/walk 4 - Randic shape index
F06[N-N]	4	0.00000	0.00034	0.00012	0.00004	2D frequency fingerprints	frequency of N - N at topological distance 06
nX	4	0.00000	0.00011	0.00011	-0.00025	constitutional descriptors	number of halogen atoms
ww	4	0.00000	0.00003	0.00011	0.00014	topological descriptors	hyper-detour index
EEig14d	4	0.00000	0.00079	0.00010	0.00036	edge adjacency indices	Eigenvalue 14 from edge adj. matrix weighted by dipole m
MATS8e	4	0.00000	0.00039	0.00010	0.00010	2D autocorrelations	Moran autocorrelation - lag 8 / weighted by atomic Sander
VEe2	4	0.00000	0.00046	0.00010	0.00039	eigenvalue-based indices	average eigenvector coefficient sum from electronegativit
F08[N-O]	4	0.00000	0.00000	0.00010	0.00033	2D frequency fingerprints	frequency of N - O at topological distance 08
GATS2v	4	0.00000	-0.00048	0.00010	0.00007	2D autocorrelations	Geary autocorrelation - lag 2 / weighted by atomic van der
F04[O-O]	4	0.00000	0.00033	0.00009	0.00052	2D frequency fingerprints	frequency of O - O at topological distance 04
F06[N-O]	4	0.00000	0.00008	0.00009	0.00026	2D frequency fingerprints	frequency of N - O at topological distance 06
F09[N-F]	4	0.00000	0.00000	0.00009	0.00006	2D frequency fingerprints	frequency of N - F at topological distance 09
F07[N-O]	4	0.00000	-0.00006	0.00008	0.00020	2D frequency fingerprints	frequency of N - O at topological distance 07
F03[N-N]	4	0.00000	0.00057	0.00007	0.00009	2D frequency fingerprints	frequency of N - N at topological distance 03
F09[N-N]	4	0.00000	0.00000	0.00007	0.00000	2D frequency fingerprints	frequency of N - N at topological distance 09
F04[O-S]	4	0.00000	0.00032	0.00006	0.00000	2D frequency fingerprints	frequency of O - S at topological distance 04
B05[O-O]	4	0.00000	0.00125	0.00006	0.00096	2D binary fingerprints	presence/absence of O - O at topological distance 05
TI2	4	0.00000	0.00086	0.00005	0.00015	topological descriptors	second Mohar index TI2
MPC10	4	0.00000	0.00063	0.00005	0.00057	walk and path counts	molecular path count of order 10
F09[N-O]	4	0.00000	0.00000	0.00004	0.00014	2D frequency fingerprints	frequency of N - O at topological distance 09
S3K	4	0.00000	0.00030	0.00002	0.00008	topological descriptors	3-path Kier alpha-modified shape index
nR07	4	0.00000	0.00037	0.00002	0.00006	constitutional descriptors	number of 7-membered rings
GATS3p	4	0.00000	-0.00031	0.00002	-0.00031	2D autocorrelations	Geary autocorrelation - lag 3 / weighted by atomic polariza
T(O..S)	4	0.00000	0.00018	0.00001	-0.00085	topological descriptors	sum of topological distances between O..S
D/Dr06	4	0.00000	0.00003	0.00001	0.00007	topological descriptors	distance/detour ring index of order 6
B03[C-C]	4	0.00000	0.00151	0.00000	0.00000	2D binary fingerprints	presence/absence of C - C at topological distance 03
B03[O-S]	4	0.00000	0.00067	0.00000	0.00000	2D binary fingerprints	presence/absence of O - S at topological distance 03
B05[S-Cl]	4	0.00000	0.00067	0.00000	0.00000	2D binary fingerprints	presence/absence of S - Cl at topological distance 05
B06[N-Cl]	4	0.00000	0.00067	0.00000	0.00000	2D binary fingerprints	presence/absence of N - Cl at topological distance 06
B07[O-S]	4	0.00000	0.00058	0.00000	0.00000	2D binary fingerprints	presence/absence of O - S at topological distance 07
B08[N-Cl]	4	0.00000	0.00000	0.00000	0.00028	2D binary fingerprints	presence/absence of N - Cl at topological distance 08
B04[O-Cl]	4	0.00000	0.00000	0.00000	0.00023	2D binary fingerprints	presence/absence of O - Cl at topological distance 04

[a] Global rank sorts the descriptors in order of decreasing global importance.

[b] Global importance measures the decrease in overall external balanced accuracy when the descriptor is permuted. Unlike the local importance score I(x,compound) which depends on the target compound, global importance G(x) measures descriptor x's contribution to the overall balanced accuracy of the model.

Table A1.3.5 (continued). Chemical descriptors and their local importance scores

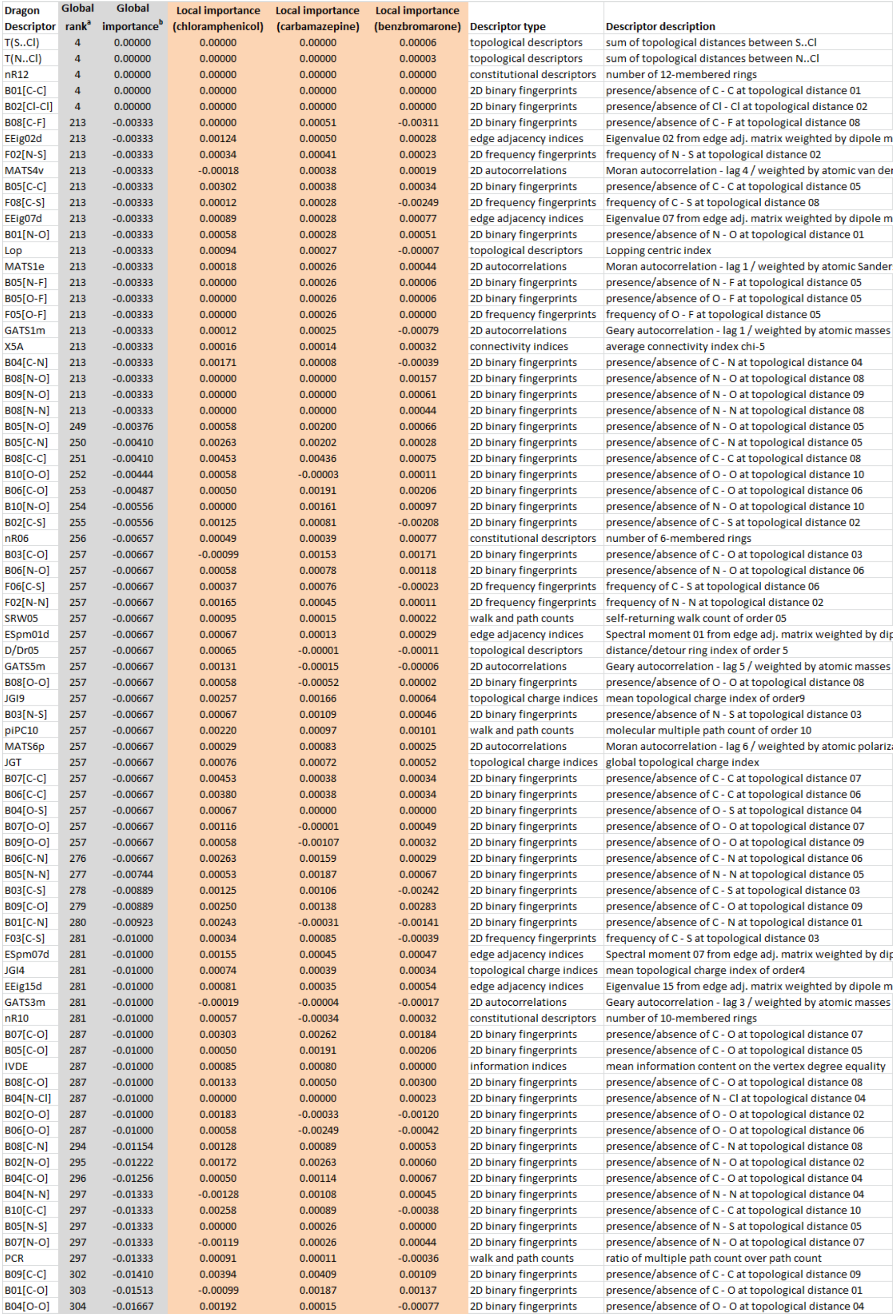


Dragon Descriptor	Global rank[a]	Global importance[b]	Local importance (chloramphenicol)	Local importance (carbamazepine)	Local importance (benzbromarone)	Descriptor type	Descriptor description
T(S..Cl)	4	0.00000	0.00000	0.00000	0.00006	topological descriptors	sum of topological distances between S..Cl
T(N..Cl)	4	0.00000	0.00000	0.00000	0.00003	topological descriptors	sum of topological distances between N..Cl
nR12	4	0.00000	0.00000	0.00000	0.00000	constitutional descriptors	number of 12-membered rings
B01[C-C]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of C - C at topological distance 01
B02[Cl-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of Cl - Cl at topological distance 02
B03[N-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of N - Cl at topological distance 03
B04[S-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of S - Cl at topological distance 04
B05[N-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of N - Cl at topological distance 05
B05[O-S]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of O - S at topological distance 05
B06[O-S]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of O - S at topological distance 06
B07[N-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of N - Cl at topological distance 07
B09[N-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of N - Cl at topological distance 09
B10[N-N]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of N - N at topological distance 10
F02[Cl-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D frequency fingerprints	frequency of Cl - Cl at topological distance 02
F09[N-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D frequency fingerprints	frequency of N - Cl at topological distance 09
B03[C-Br]	4	0.00000	0.00000	0.00000	-0.00048	2D binary fingerprints	presence/absence of C - Br at topological distance 03
F04[C-Br]	4	0.00000	0.00000	0.00000	-0.00048	2D frequency fingerprints	frequency of C - Br at topological distance 04
nI	4	0.00000	0.00000	0.00000	-0.00068	constitutional descriptors	number of Iodine atoms
T(S..F)	4	0.00000	0.00000	0.00000	-0.00211	topological descriptors	sum of topological distances between S..F
nCIR	4	0.00000	0.00026	0.00000	0.00013	constitutional descriptors	number of circuits
GATS3e	4	0.00000	0.00025	-0.00001	-0.00058	2D autocorrelations	Geary autocorrelation - lag 3 / weighted by atomic Sanders
X3A	4	0.00000	0.00087	-0.00002	0.00004	connectivity indices	average connectivity index chi-3
MSD	4	0.00000	0.00010	-0.00007	0.00039	topological descriptors	mean square distance index (Balaban)
nR09	4	0.00000	0.00048	-0.00008	0.00016	constitutional descriptors	number of 9-membered rings
GATS4m	4	0.00000	0.00017	-0.00009	-0.00009	2D autocorrelations	Geary autocorrelation - lag 4 / weighted by atomic masses
B04[C-Cl]	4	0.00000	0.00133	-0.00009	0.00052	2D binary fingerprints	presence/absence of C - Cl at topological distance 04
GATS2p	4	0.00000	-0.00065	-0.00016	-0.00023	2D autocorrelations	Geary autocorrelation - lag 2 / weighted by atomic polariza
F07[O-Cl]	4	0.00000	0.00017	-0.00019	0.00006	2D frequency fingerprints	frequency of O - Cl at topological distance 07
D/Dr09	4	0.00000	0.00093	-0.00027	-0.00034	topological descriptors	distance/detour ring index of order 9
F07[C-Cl]	4	0.00000	0.00017	-0.00032	0.00008	2D frequency fingerprints	frequency of C - Cl at topological distance 07
JGI5	4	0.00000	0.00126	-0.00038	0.00027	topological charge indices	mean topological charge index of order5
B03[C-Cl]	4	0.00000	0.00067	-0.00038	0.00051	2D binary fingerprints	presence/absence of C - Cl at topological distance 03
B06[C-Cl]	4	0.00000	0.00000	-0.00038	0.00051	2D binary fingerprints	presence/absence of C - Cl at topological distance 06
B10[C-Cl]	4	0.00000	0.00067	-0.00057	0.00028	2D binary fingerprints	presence/absence of C - Cl at topological distance 10
F06[C-Cl]	4	0.00000	0.00000	-0.00059	0.00014	2D frequency fingerprints	frequency of C - Cl at topological distance 06
F08[C-Cl]	4	0.00000	0.00013	-0.00061	0.00000	2D frequency fingerprints	frequency of C - Cl at topological distance 08
F06[O-Cl]	4	0.00000	0.00034	-0.00063	0.00023	2D frequency fingerprints	frequency of O - Cl at topological distance 06
F04[C-Cl]	4	0.00000	0.00067	-0.00073	0.00027	2D frequency fingerprints	frequency of C - Cl at topological distance 04
B07[O-Cl]	4	0.00000	0.00067	-0.00077	0.00000	2D binary fingerprints	presence/absence of O - Cl at topological distance 07
B09[O-Cl]	4	0.00000	0.00067	-0.00077	0.00000	2D binary fingerprints	presence/absence of O - Cl at topological distance 09
B03[O-Cl]	4	0.00000	0.00000	-0.00078	0.00000	2D binary fingerprints	presence/absence of O - Cl at topological distance 03
F08[O-Cl]	4	0.00000	0.00034	-0.00080	0.00000	2D frequency fingerprints	frequency of O - Cl at topological distance 08
F05[C-Cl]	4	0.00000	0.00045	-0.00081	0.00004	2D frequency fingerprints	frequency of C - Cl at topological distance 05
B09[C-Cl]	4	0.00000	0.00067	-0.00142	0.00028	2D binary fingerprints	presence/absence of C - Cl at topological distance 09
B05[C-Cl]	4	0.00000	0.00000	-0.00150	0.00051	2D binary fingerprints	presence/absence of C - Cl at topological distance 05
B08[O-Cl]	4	0.00000	0.00133	-0.00238	0.00000	2D binary fingerprints	presence/absence of O - Cl at topological distance 08
B05[O-Cl]	4	0.00000	0.00067	-0.00238	0.00000	2D binary fingerprints	presence/absence of O - Cl at topological distance 05
B08[C-Cl]	4	0.00000	0.00067	-0.00280	0.00023	2D binary fingerprints	presence/absence of C - Cl at topological distance 08
B06[O-Cl]	4	0.00000	0.00067	-0.00280	0.00022	2D binary fingerprints	presence/absence of O - Cl at topological distance 06
B10[O-Cl]	4	0.00000	0.00000	-0.00382	0.00000	2D binary fingerprints	presence/absence of O - Cl at topological distance 10
nR11	203	-0.00010	0.00074	0.00014	0.00007	constitutional descriptors	number of 11-membered rings
B02[N-N]	204	-0.00077	0.00239	0.00211	0.00051	2D binary fingerprints	presence/absence of N - N at topological distance 02
B01[N-N]	205	-0.00077	0.00125	0.00025	0.00102	2D binary fingerprints	presence/absence of N - N at topological distance 01
B03[C-N]	206	-0.00179	0.00321	0.00133	0.00055	2D binary fingerprints	presence/absence of C - N at topological distance 03
nTB	207	-0.00215	0.00000	0.00020	-0.00047	constitutional descriptors	number of triple bonds
B10[C-N]	208	-0.00222	0.00000	0.00212	0.00199	2D binary fingerprints	presence/absence of C - N at topological distance 10
B09[C-N]	208	-0.00222	0.00125	0.00186	0.00130	2D binary fingerprints	presence/absence of C - N at topological distance 09
B10[C-O]	208	-0.00222	0.00250	0.00162	0.00232	2D binary fingerprints	presence/absence of C - O at topological distance 10
B05[C-S]	208	-0.00222	0.00125	0.00106	-0.00208	2D binary fingerprints	presence/absence of C - S at topological distance 05
nF	212	-0.00323	0.00000	0.00026	-0.00101	constitutional descriptors	number of Fluorine atoms
ATS8m	213	-0.00333	0.00205	0.00118	0.00056	2D autocorrelations	Broto-Moreau autocorrelation of a topological structure - l
MATS4e	213	-0.00333	-0.00005	0.00062	0.00029	2D autocorrelations	Moran autocorrelation - lag 4 / weighted by atomic Sander
JGI10	213	-0.00333	0.00158	0.00061	0.00094	topological charge indices	mean topological charge index of order10
F05[C-S]	213	-0.00333	0.00063	0.00055	-0.00017	2D frequency fingerprints	frequency of C - S at topological distance 05
GATS5p	213	-0.00333	0.00135	0.00024	0.00056	2D autocorrelations	Geary autocorrelation - lag 5 / weighted by atomic polariza
B03[O-O]	213	-0.00333	0.00067	0.00018	0.00090	2D binary fingerprints	presence/absence of O - O at topological distance 03
SRW09	213	-0.00333	0.00081	0.00002	0.00014	walk and path counts	self-returning walk count of order 09
MATS7e	213	-0.00333	0.00019	0.00100	-0.00038	2D autocorrelations	Moran autocorrelation - lag 7 / weighted by atomic Sander
JGI2	213	-0.00333	0.00112	0.00085	-0.00015	topological charge indices	mean topological charge index of order2
B07[C-S]	213	-0.00333	0.00058	0.00081	-0.00208	2D binary fingerprints	presence/absence of C - S at topological distance 07
B04[C-C]	213	-0.00333	0.00000	0.00077	0.00102	2D binary fingerprints	presence/absence of C - C at topological distance 04
GATS6v	213	-0.00333	0.00110	0.00076	0.00007	2D autocorrelations	Geary autocorrelation - lag 6 / weighted by atomic van der
GATS6p	213	-0.00333	0.00094	0.00066	0.00014	2D autocorrelations	Geary autocorrelation - lag 6 / weighted by atomic polariza
MATS3m	213	-0.00333	0.00147	0.00061	0.00044	2D autocorrelations	Moran autocorrelation - lag 3 / weighted by atomic masses
MATS6m	213	-0.00333	0.00039	0.00061	0.00011	2D autocorrelations	Moran autocorrelation - lag 6 / weighted by atomic masses
B02[N-S]	213	-0.00333	0.00067	0.00053	0.00023	2D binary fingerprints	presence/absence of N - S at topological distance 02
B02[C-F]	213	-0.00333	0.00000	0.00051	-0.00311	2D binary fingerprints	presence/absence of C - F at topological distance 02

[a] Global rank sorts the descriptors in order of decreasing global importance.

[b] Global importance measures the decrease in overall external balanced accuracy when the descriptor is permuted. Unlike the local importance score I(x,compound) which depends on the target compound, global importance G(x) measures descriptor x's contribution to the overall balanced accuracy of the model.

Table A1.3.5 (continued). Chemical descriptors and their local importance scores



Dragon Descriptor	Global rank[a]	Global importance[b]	Local importance (chloramphenicol)	Local importance (carbamazepine)	Local importance (benzbromarone)	Descriptor type	Descriptor description
T(S..Cl)	4	0.00000	0.00000	0.00000	0.00006	topological descriptors	sum of topological distances between S..Cl
T(N..Cl)	4	0.00000	0.00000	0.00000	0.00003	topological descriptors	sum of topological distances between N..Cl
nR12	4	0.00000	0.00000	0.00000	0.00000	constitutional descriptors	number of 12-membered rings
B01[C-C]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of C - C at topological distance 01
B02[Cl-Cl]	4	0.00000	0.00000	0.00000	0.00000	2D binary fingerprints	presence/absence of Cl - Cl at topological distance 02
B08[C-F]	213	-0.00333	0.00000	0.00051	-0.00311	2D binary fingerprints	presence/absence of C - F at topological distance 08
EEig02d	213	-0.00333	0.00124	0.00050	0.00028	edge adjacency indices	Eigenvalue 02 from edge adj. matrix weighted by dipole m
F02[N-S]	213	-0.00333	0.00034	0.00041	0.00023	2D frequency fingerprints	frequency of N - S at topological distance 02
MATS4v	213	-0.00333	-0.00018	0.00038	0.00019	2D autocorrelations	Moran autocorrelation - lag 4 / weighted by atomic van der
B05[C-C]	213	-0.00333	0.00302	0.00038	0.00034	2D binary fingerprints	presence/absence of C - C at topological distance 05
F08[C-S]	213	-0.00333	0.00012	0.00028	-0.00249	2D frequency fingerprints	frequency of C - S at topological distance 08
EEig07d	213	-0.00333	0.00089	0.00028	0.00077	edge adjacency indices	Eigenvalue 07 from edge adj. matrix weighted by dipole m
B01[N-O]	213	-0.00333	0.00058	0.00028	0.00051	2D binary fingerprints	presence/absence of N - O at topological distance 01
Lop	213	-0.00333	0.00094	0.00027	-0.00007	topological descriptors	Lopping centric index
MATS1e	213	-0.00333	0.00018	0.00026	0.00044	2D autocorrelations	Moran autocorrelation - lag 1 / weighted by atomic Sander
B05[N-F]	213	-0.00333	0.00000	0.00026	0.00006	2D binary fingerprints	presence/absence of N - F at topological distance 05
B05[O-F]	213	-0.00333	0.00000	0.00026	0.00006	2D binary fingerprints	presence/absence of O - F at topological distance 05
F05[O-F]	213	-0.00333	0.00000	0.00026	0.00000	2D frequency fingerprints	frequency of O - F at topological distance 05
GATS1m	213	-0.00333	0.00012	0.00025	-0.00079	2D autocorrelations	Geary autocorrelation - lag 1 / weighted by atomic masses
X5A	213	-0.00333	0.00016	0.00014	0.00032	connectivity indices	average connectivity index chi-5
B04[C-N]	213	-0.00333	0.00171	0.00008	-0.00039	2D binary fingerprints	presence/absence of C - N at topological distance 04
B08[N-O]	213	-0.00333	0.00000	0.00000	0.00157	2D binary fingerprints	presence/absence of N - O at topological distance 08
B09[N-O]	213	-0.00333	0.00000	0.00000	0.00061	2D binary fingerprints	presence/absence of N - O at topological distance 09
B08[N-N]	213	-0.00333	0.00000	0.00000	0.00044	2D binary fingerprints	presence/absence of N - N at topological distance 08
B05[N-O]	249	-0.00376	0.00058	0.00200	0.00066	2D binary fingerprints	presence/absence of N - O at topological distance 05
B05[C-N]	250	-0.00410	0.00263	0.00202	0.00028	2D binary fingerprints	presence/absence of C - N at topological distance 05
B08[C-C]	251	-0.00410	0.00453	0.00436	0.00075	2D binary fingerprints	presence/absence of C - C at topological distance 08
B10[O-O]	252	-0.00444	0.00058	-0.00003	0.00011	2D binary fingerprints	presence/absence of O - O at topological distance 10
B06[C-O]	253	-0.00487	0.00050	0.00191	0.00206	2D binary fingerprints	presence/absence of C - O at topological distance 06
B10[N-O]	254	-0.00556	0.00000	0.00161	0.00097	2D binary fingerprints	presence/absence of N - O at topological distance 10
B02[C-S]	255	-0.00556	0.00125	0.00081	-0.00208	2D binary fingerprints	presence/absence of C - S at topological distance 02
nR06	256	-0.00657	0.00049	0.00039	0.00077	constitutional descriptors	number of 6-membered rings
B03[C-O]	257	-0.00667	-0.00099	0.00153	0.00171	2D binary fingerprints	presence/absence of C - O at topological distance 03
B06[N-O]	257	-0.00667	0.00058	0.00078	0.00118	2D binary fingerprints	presence/absence of N - O at topological distance 06
F06[C-S]	257	-0.00667	0.00037	0.00076	-0.00023	2D frequency fingerprints	frequency of C - S at topological distance 06
F02[N-N]	257	-0.00667	0.00165	0.00045	0.00011	2D frequency fingerprints	frequency of N - N at topological distance 02
SRW05	257	-0.00667	0.00095	0.00015	0.00022	walk and path counts	self-returning walk count of order 05
ESpm01d	257	-0.00667	0.00067	0.00013	0.00029	edge adjacency indices	Spectral moment 01 from edge adj. matrix weighted by dip
D/Dr05	257	-0.00667	0.00065	-0.00001	-0.00011	topological descriptors	distance/detour ring index of order 5
GATS5m	257	-0.00667	0.00131	-0.00015	-0.00006	2D autocorrelations	Geary autocorrelation - lag 5 / weighted by atomic masses
B08[O-O]	257	-0.00667	0.00058	-0.00052	0.00002	2D binary fingerprints	presence/absence of O - O at topological distance 08
JGI9	257	-0.00667	0.00257	0.00166	0.00064	topological charge indices	mean topological charge index of order9
B03[N-S]	257	-0.00667	0.00067	0.00109	0.00046	2D binary fingerprints	presence/absence of N - S at topological distance 03
piPC10	257	-0.00667	0.00220	0.00097	0.00101	walk and path counts	molecular multiple path count of order 10
MATS6p	257	-0.00667	0.00029	0.00083	0.00025	2D autocorrelations	Moran autocorrelation - lag 6 / weighted by atomic polariz
JGT	257	-0.00667	0.00076	0.00072	0.00052	topological charge indices	global topological charge index
B07[C-C]	257	-0.00667	0.00453	0.00038	0.00034	2D binary fingerprints	presence/absence of C - C at topological distance 07
B06[C-C]	257	-0.00667	0.00380	0.00038	0.00034	2D binary fingerprints	presence/absence of C - C at topological distance 06
B04[O-S]	257	-0.00667	0.00067	0.00000	0.00000	2D binary fingerprints	presence/absence of O - S at topological distance 04
B07[O-O]	257	-0.00667	0.00116	-0.00001	0.00049	2D binary fingerprints	presence/absence of O - O at topological distance 07
B09[O-O]	257	-0.00667	0.00058	-0.00107	0.00032	2D binary fingerprints	presence/absence of O - O at topological distance 09
B06[C-N]	276	-0.00667	0.00263	0.00159	0.00029	2D binary fingerprints	presence/absence of C - N at topological distance 06
B05[N-N]	277	-0.00744	0.00053	0.00187	0.00067	2D binary fingerprints	presence/absence of N - N at topological distance 05
B03[C-S]	278	-0.00889	0.00125	0.00106	-0.00242	2D binary fingerprints	presence/absence of C - S at topological distance 03
B09[C-O]	279	-0.00889	0.00250	0.00138	0.00283	2D binary fingerprints	presence/absence of C - O at topological distance 09
B01[C-N]	280	-0.00923	0.00243	-0.00031	-0.00141	2D binary fingerprints	presence/absence of C - N at topological distance 01
F03[C-S]	281	-0.01000	0.00034	0.00085	-0.00039	2D frequency fingerprints	frequency of C - S at topological distance 03
ESpm07d	281	-0.01000	0.00155	0.00045	0.00047	edge adjacency indices	Spectral moment 07 from edge adj. matrix weighted by dip
JGI4	281	-0.01000	0.00074	0.00039	0.00034	topological charge indices	mean topological charge index of order4
EEig15d	281	-0.01000	0.00081	0.00035	0.00054	edge adjacency indices	Eigenvalue 15 from edge adj. matrix weighted by dipole m
GATS3m	281	-0.01000	-0.00019	-0.00004	-0.00017	2D autocorrelations	Geary autocorrelation - lag 3 / weighted by atomic masses
nR10	281	-0.01000	0.00057	-0.00034	0.00032	constitutional descriptors	number of 10-membered rings
B07[C-O]	287	-0.01000	0.00303	0.00262	0.00184	2D binary fingerprints	presence/absence of C - O at topological distance 07
B05[C-O]	287	-0.01000	0.00050	0.00191	0.00206	2D binary fingerprints	presence/absence of C - O at topological distance 05
IVDE	287	-0.01000	0.00085	0.00080	0.00000	information indices	mean information content on the vertex degree equality
B08[C-O]	287	-0.01000	0.00133	0.00050	0.00300	2D binary fingerprints	presence/absence of C - O at topological distance 08
B04[N-Cl]	287	-0.01000	0.00000	0.00000	0.00023	2D binary fingerprints	presence/absence of N - Cl at topological distance 04
B02[O-O]	287	-0.01000	0.00183	-0.00033	-0.00120	2D binary fingerprints	presence/absence of O - O at topological distance 02
B06[O-O]	287	-0.01000	0.00058	-0.00249	-0.00042	2D binary fingerprints	presence/absence of O - O at topological distance 06
B08[C-N]	294	-0.01154	0.00128	0.00089	0.00053	2D binary fingerprints	presence/absence of C - N at topological distance 08
B02[N-O]	295	-0.01222	0.00172	0.00263	0.00060	2D binary fingerprints	presence/absence of N - O at topological distance 02
B04[C-O]	296	-0.01256	0.00050	0.00114	0.00067	2D binary fingerprints	presence/absence of C - O at topological distance 04
B04[N-N]	297	-0.01333	-0.00128	0.00108	0.00045	2D binary fingerprints	presence/absence of N - N at topological distance 04
B10[C-C]	297	-0.01333	0.00258	0.00089	-0.00038	2D binary fingerprints	presence/absence of C - C at topological distance 10
B05[N-S]	297	-0.01333	0.00000	0.00026	0.00000	2D binary fingerprints	presence/absence of N - S at topological distance 05
B07[N-O]	297	-0.01333	-0.00119	0.00026	0.00044	2D binary fingerprints	presence/absence of N - O at topological distance 07
PCR	297	-0.01333	0.00091	0.00011	-0.00036	walk and path counts	ratio of multiple path count over path count
B09[C-C]	302	-0.01410	0.00394	0.00409	0.00109	2D binary fingerprints	presence/absence of C - C at topological distance 09
B01[C-O]	303	-0.01513	-0.00099	0.00187	0.00137	2D binary fingerprints	presence/absence of C - O at topological distance 01
B04[O-O]	304	-0.01667	0.00192	0.00015	-0.00077	2D binary fingerprints	presence/absence of O - O at topological distance 04

[a] Global rank sorts the descriptors in order of decreasing global importance.

[b] Global importance measures the decrease in overall external balanced accuracy when the descriptor is permuted. Unlike the local importance score I(x,compound) which depends on the target compound, global importance G(x) measures descriptor x's contribution to the overall balanced accuracy of the model.

Supplemental tables for Chapter 4

Table A1.4.S1. Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
armodafinil	1	N06	NC(=O)CS(=O)C(c1:c:c:c:c:c:1)c2:c:c:c:c:c:2
levofloxacin	1	J01,S01	CC1COc2:c(N3CCN(C)CC3):c(F):c:c4C(=O)C(=CN1c:2:4)C(=O)O
abacavir	1	J05	Nc1:n:c(NC2CC2):c3:n:c:n(C4CC(CO)C=C4):c:3:n:1
acetylcysteine	1	R05,S01,V03	CC(=O)NC(CS)C(=O)O
acetylsalicylic_acid	1	A01,B01,N02	CC(=O)Oc1:c:c:c:c:c:1C(=O)O
aciclovir	1	D06,J05,S01	Nc1:n:c(O):c2:n:c:n(COCCO):c:2:n:1
albendazole	1	P02	CCCSc1:c:c:c2:[nH]:c(NC(=O)OC):n:c:2:c:1
allopurinol	1	M04	Oc1:n:c:n:c2:n:[nH]:c:c:1:2
amantadine	1	N04	NC12CC3CC(CC(C3)C1)C2
ambroxol	1	R05	Nc1:c(Br):c:c(Br):c:c:1CNC2CCC(O)CC2
amifostine	1	V03	NCCCNCCSP(=O)(O)O
aminoglutethimide	1	L02	CCC1(CCC(=O)NC1=O)c2:c:c:c(N):c:c:2
amoxicillin	1	J01	CC1(C)SC2C(NC(=O)C(N)c3:c:c:c(O):c:c:3)C(=O)N2C1C(=O)O
ampicillin	1	J01,S01	CC1(C)SC2C(NC(=O)C(N)c3:c:c:c:c:c:3)C(=O)N2C1C(=O)O
ascorbic_acid	1	G01,S01	OCC(O)C1OC(=O)C(O)C1=O
azithromycin	1	J01,S01	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)(O)CC(C)CN(C)C(C)C(O)C1(C)O
bacampicillin	1	J01	CCOC(=O)OC(C)OC(=O)C1N2C(SC1(C)C)C(NC(=O)C(N)c3:c:c:c:c:c:3)C2=O
bendamustine	1	L01	Cn1:c(CCCC(=O)O):n:c2:c:c(:c:c:c:1:2)N(CCCl)CCCl
benoxaprofen	1	M01	CC(C(=O)O)c1:c:c:c2:o:c(:n:c:2:c:1)c3:c:c:c(Cl):c:c:3
benzylpenicillin	1	J01,S01	CC1(C)SC2C(NC(=O)Cc3:c:c:c:c:c:3)C(=O)N2C1C(=O)O
beta-acetyldigoxin	1	C01	CC1OC(CC(O)C1OC2CC(O)C(OC3CC(O)C(OC(=O)C)C(C)O3)C(C)O2)OC4CCC5(C)C(CCC6C5CC(O)C7(C)C(CCC67O)C8=CC(=O)OC8)C4
bisacodyl	1	A06	CC(=O)Oc1:c:c:c(:c:c:1)C(c2:c:c:c(OC(=O)C):c:c:2)c3:c:c:c:c:n:3
bisoprolol	1	C07	CC(C)NCC(O)COc1:c:c:c(COCCOC(C)C):c:c:1
bromhexine	1	R05	CN(Cc1:c:c(Br):c:c(Br):c:1N)C2CCCCC2
brotizolam	1	N05	Cc1:n:n:c2CN=C(c3:c:c:c:c:c:3Cl)c4:c:c(Br):s:c:4n:1:2
bupropion	1	N06	CC(NC(C)(C)C)C(=O)c1:c:c:c:c(Cl):c:1
carbamazepine	1	N03	NC(=O)N1c2:c:c:c:c:c:2C=Cc3:c:c:c:c:c1:3
carbimazole	1	H03	CCOC(=O)N1C=CN(C)C1=S
carbocisteine	1	R05	NC(CSCC(=O)O)C(=O)O
cefaclor	1	J01	NC(C(=O)NC1C2SCC(=C(N2C1=O)C(=O)O)Cl)c3:c:c:c:c:c:3
cefadroxil	1	J01	CC1=C(N2C(SC1)C(NC(=O)C(N)c3:c:c:c(O):c:c:3)C2=O)C(=O)O
cefalexin	1	J01	CC1=C(N2C(SC1)C(NC(=O)C(N)c3:c:c:c:c:c:3)C2=O)C(=O)O
cefazolin	1	J01	Cc1:n:n:c(SCC2=C(N3C(SC2)C(NC(=O)Cn4:c:n:n:n:4)C3=O)C(=O)O):s:1
cefdinir	1	J01	Nc1:n:c(:c:s:1)C(N=O)C(=O)NC2C3SCC(=C(N3C2=O)C(=O)O)C=C
cefepime	1	J01	CO\N=C(\C(=O)NC1C2SCC(=C(N2C1=O)C(=O)O)CN3(C)CCCC3)/c4:c:s:c(N):n:4
cefixime	1	J01	Nc1:n:c(:c:s:1)\C(=N/OCC(=O)O)\C(=O)NC2C3SCC(=C(N3C2=O)C(=O)O)C=C
cefotiam	1	J01	CN(C)CCn1:n:n:n:c:1SCC2=C(N3C(SC2)C(NC(=O)Cc4:c:s:c(N):n:4)C3=O)C(=O)O
cefpodoxime	1	J01	COCC1=C(N2C(SC1)C(NC(=O)\C(=N\OC)\c3:c:s:c(N):n:3)C2=O)C(=O)O
cefprozil	1	J01	C\C=C\C1=C(N2C(SC1)C(NC(=O)C(N)c3:c:c:c(O):c:c:3)C2=O)C(=O)O
ceftazidime	1	J01	CC(C)(O\N=C(\C(=N\C1C2SCC(=C(N2C1=O)C(=O)O)CN3=CC=CC=C3)\O)/C4=CSC(=N)N4)C(=O)O
ceftibuten	1	J01	Nc1:n:c(:c:s:1)\C(=C/CC(=O)O)\C(=O)NC2C3SCC=C(N3C2=O)C(=O)O
ceftriaxone	1	J01	CO\N=C(\C(=O)NC1C2SCC(=C(N2C1=O)C(=O)O)CSC3=NC(=O)C(=O)NN3C)/c4:c:s:c(N):n:4
cefuroxime	1	J01	CO\N=C(\C(=O)NC1C2SCC(=C(N2C1=O)C(=O)O)COC(=O)N)/c3:o:c:c:c:3
chloral_hydrate	1	N05	OC(O)C(Cl)(Cl)Cl
chloramphenicol	1	D06,D10,G01,J01,S01,S02,S03	OCC(NC(=O)C(Cl)Cl)C(O)c1:c:c:c(:c:c:1)N(=O)=O
chlormezanone	1	M03	CN1C(c2:c:c:c(Cl):c:c:2)S(=O)(=O)CCC1=O
chloroquine	1	P01	CCN(CC)CCCC(C)Nc1:c:c:n:c2:c:c(Cl):c:c:c:1:2
chlorphenamine	1	R06	CN(C)CCC(c1:c:c:c(Cl):c:c:1)c2:c:c:c:c:n:2
chlorpropamide	1	A10	CCCNC(=O)NS(=O)(=O)c1:c:c:c(Cl):c:c:1
chlortetracycline	1	A01,D06,J01,S01	CN(C)C1C2CC3C(C(=O)c4:c(O):c:c:c(Cl):c:4C3(C)O)C(=O)C2(O)C(=O)C(C(=O)N)C1=O
cinnarizine	1	N07	C(\C=C\c1:c:c:c:c:c:1)N2CCN(CC2)C(c3:c:c:c:c:c:3)c4:c:c:c:c:c:4
ciprofloxacin	1	J01,S01,S02,S03	OC(=O)C1=CN(C2CC2)c3:c:c(N4CCNCC4):c(F):c:c:3C1=O
clarithromycin	1	J01	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)(CC(C)C(=O)C(C)C(O)C1(C)O)OC
clemastine	1	D04,R06	CN1CCCC1CCOC(C)(c2:c:c:c:c:c:2)c3:c:c:c(Cl):c:c:3
clindamycin	1	D10,G01,J01	CCCC1CC(N(C)C1)C(=O)NC(C(C)Cl)C2OC(SC)C(O)C(O)C2O
clobutinol	1	R05	CC(CN(C)C)C(C)(O)Cc1:c:c:c(Cl):c:c:1
clomethiazole	1	N05	Cc1:n:c:s:c:1CCCl
clotrimazole	1	A01,D01,G01	Clc1:c:c:c:c:c:1C(c2:c:c:c:c:c:2)(c3:c:c:c:c:c:3)n4:c:c:n:c:4
cloxacillin	1	J01	Cc1:o:n:c(c2:c:c:c:c:c:2Cl):c:1C(=O)NC3C4SC(C)(C)C(N4C3=O)C(=O)O

Table A1.4.S1 (continued). Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
codeine	1	R05	COc1:c:c:c2CC3C4C=CC(O)C5Oc:1:c:2C45CCN3C
colchicine	1	M04	COC1=CC=C2C(=CC1=O)C(CCc3:c:c(OC):c(OC):c(OC):c2:3)NC(=O)C
dapsone	1	D10,J04	Nc1:c:c:c(:c:c:1)S(=O)(=O)c2:c:c:c(N):c:c:2
darunavir	1	J05	CC(C)CN(CC(O)C(Cc1:c:c:c:c:c:1)NC(=O)OC2COC3OCCC23)S(=O)(=O)c4:c:c:c(N):c:c:4
dexamethasone	1	A01,C05,D07,D10,H02,R01,S01,S02,S03	CC1CC2C3CCC4=CC(=O)C=CC4(C)C3(F)C(O)CC2(C)C1(O)C(=O)CO
dexpanthenol	1	A11,D03,S01	CC(C)(CO)C(O)C(=O)NCCCO
diclofenac	1	D11,M01,M02,S01	OC(=O)Cc1:c:c:c:c:c:1Nc2:c(Cl):c:c:c:c:2Cl
dicloxacillin	1	J01	Cc1:o:n:c(:c:1C(=O)NC2C3SC(C)(C)C(N3C2=O)C(=O)O)c4:c(Cl):c:c:c:c:4Cl
diflunisal	1	N02	OC(=O)c1:c:c(:c:c:c:1O)c2:c:c:c(F):c:c:2F
digitoxin	1	C01	CC1OC(CC(O)C1O)OC2C(O)CC(OC3C(O)CC(OC4CCC5(C)C(CCC6C5CCC7(C)C(CCC67O)C8=CC(=O)OC8)C4)OC3C)OC2C
dihydrocodeine	1	N02	COc1:c:c:c2CC3C4CCC(O)C5Oc:1:c:2C45CCN3C
diltiazem	1	C08	COc1:c:c:c(:c:c:1)C2Sc3:c:c:c:c:c:3N(CCN(C)C)C(=O)C2OC(=O)C
dimetindene	1	D04,R06	CC(C1=C(CCN(C)C)Cc2:c:c:c:c:c1:2)c3:c:c:c:c:n:3
doxycycline	1	A01,J01	CC1[C@H]2[C@@H](O)[C@H]3[C@@H](N(C)C)C(=O)C(C(=O)N)C(=O)[C@]3(O)C(=O)C2C(=O)c4:c(O):c:c:c:c1:4
efavirenz	1	J05	FC(F)(F)C1(OC(=O)Nc2:c:c:c(Cl):c:c1:2)C#CC3CC3
erythromycin	1	D10,J01,S01	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)(O)CC(C)C(=O)C(C)C(O)C1(C)O
ethambutol	1	J04	CCC(CO)NCCNC(CC)CO
etodolac	1	M01	CCc1:c:c:c:c2:c3CCOC(CC)(CC(=O)O)c:3:[nH]:c:1:2
etoricoxib	1	M01	Cc1:c:c:c(:c:n:1)c2:n:c:c(Cl):c:c:2c3:c:c:c(:c:c:3)S(=O)(=O)C
etravirine	1	J05	Cc1:c:c(:c:c(C):c:1Oc2:n:c(Nc3:c:c:c(:c:c:3)C#N):n:c(N):c:2Br)C#N
felbamate	1	N03	NC(=O)OCC(COC(=O)N)c1:c:c:c:c:c:1
fenbufen	1	M01	OC(=O)CCC(=O)c1:c:c:c(:c:c:1)c2:c:c:c:c:c:2
feprazone	1	M01,M02	CC(=CCC1C(=O)N(N(C1=O)c2:c:c:c:c:c:2)c3:c:c:c:c:c:3)C
fluconazole	1	D01,J02	OC(Cn1:c:n:c:n:1)(Cn2:c:n:c:n:2)c3:c:c:c(F):c:c:3F
fosphenytoin	1	N03	OP(=O)(O)OCN1C(=O)NC(C1=O)(c2:c:c:c:c:c:2)c3:c:c:c:c:c:3
furosemide	1	C03	NS(=O)(=O)c1:c:c(C(=O)O):c(NCc2:o:c:c:c:2):c:c:1Cl
gatifloxacin	1	J01,S01	COc1:c(N2CCNC(C)C2):c(F):c:c3C(=O)C(=CN(C4CC4)c:1:3)C(=O)O
gemifloxacin	1	J01	CO\N=C\1/CN(CC1CN)c2:n:c3N(C=C(C(=O)O)C(=O)c:3:c:c:2F)C4CC4
glibenclamide	1	A10	COc1:c:c:c(Cl):c:c:1C(=O)NCCc2:c:c:c(:c:c:2)S(=O)(=O)NC(=O)NC3CCCCC3
glimepiride	1	A10	CCC1=C(C)CN(C(=O)NCCc2:c:c:c(:c:c:2)S(=O)(=O)NC(=O)NC3CCC(C)CC3)C1=O
glipizide	1	A10	Cc1:c:n:c(:c:n:1)C(=O)NCCc2:c:c:c(:c:c:2)S(=O)(=O)NC(=O)NC3CCCCC3
griseofulvin	1	D01	COC1=CC(=O)CC(C)C12Oc3:c(Cl):c(OC):c:c(OC):c:3C2=O
hexetidine	1	A01	CCCCC(CC)CN1CN(CC(CC)CCCC)CC(C)(N)C1
hydrochlorothiazide	1	C03	NS(=O)(=O)c1:c:c2:c(NCNS2(=O)=O):c:c:1Cl
hydroxychloroquine	1	P01	CCN(CCO)CCCC(C)Nc1:c:c:n:c2:c:c(Cl):c:c:c:1:2
ibuprofen	1	C01,G02,M01,M02	CC(C)Cc1:c:c:c(:c:c:1)C(C)C(=O)O
indapamide	1	C03	CC1Cc2:c:c:c:c:c:2N1NC(=O)c3:c:c:c(Cl):c(:c:3)S(=O)(=O)N
indometacin	1	C01,M01,M02,S01	COc1:c:c:c2:c(:c:1):c(CC(=O)O):c(C):n:2C(=O)c3:c:c:c(Cl):c:c:3
isoniazid	1	J04	NNC(=O)c1:c:c:n:c:c:1
isopromethazine	1	R06	CC(CN1c2:c:c:c:c:c:2Sc3:c:c:c:c:c1:3)N(C)C
isosorbide_mononitrate	1	C01	OC1COC2C(COC12)ON(=O)=O
itraconazole	1	J02	CCC(C)N1N=CN(C1=O)c2:c:c:c(:c:c:2)N3CCN(CC3)c4:c:c:c(OCC5COC(Cn6:c:n:c:n:6)(O5)c7:c:c:c(Cl):c:c:7Cl):c:c:4
josamycin	1	J01	COC1C(CC(=O)OC(C)C\C=C/C=C/C(O)C(C)CC(CC=O)C1OC2OC(C)C(OC3CC(C)(O)C(OC(=O)CC(C)C)C(C)O3)C(C2O)N(C)C)OC(=O)C
kanamycin	1	A07,J01,S01	NCC1OC(OC2C(N)CC(N)C(OC3OC(CO)C(O)C(N)C3O)C2O)C(O)C(O)C1O
ketoconazole	1	D01,G01,J02	CC(=O)N1CCN(CC1)c2:c:c:c(OCC3COC(Cn4:c:c:n:c:4)(O3)c5:c:c:c(Cl):c:c:5Cl):c:c:2
lamotrigine	1	N03	Nc1:n:n:c(:c(N):n:1)c2:c:c:c:c(Cl):c:2Cl
leflunomide	1	L04	Cc1:o:n:c:c:1C(=O)Nc2:c:c:c(:c:c:2)C(F)(F)F
lenalidomide	1	L04	Nc1:c:c:c:c2C(=O)N(Cc:1:2)C3CCC(=O)NC3=O
lincomycin	1	J01	CCCC1CC(N(C)C1)C(=O)NC(C(C)O)C2OC(SC)C(O)C(O)C2O
linezolid	1	J01	CC(=O)NCC1CN(C(=O)O1)c2:c:c:c(N3CCOCC3):c(F):c:2
loracarbef	1	J01	NC(C(=O)NC1C2CCC(=C(N2C1=O)C(=O)O)Cl)c3:c:c:c:c:c:3
loxoprofen	1	M01,M02	CC(C(=O)O)c1:c:c:c(CC2CCCC2=O):c:c:1
lymecycline	1	J01	CN(C)C1C2CC3C(C(=O)c4:c(O):c:c:c:c:4C3(C)O)C(=O)C2(O)C(=O)C(C(=O)NCNCCCCC(N)C(=O)O)C1=O
maprotiline	1	N06	CNCCCC12CCC(c3:c:c:c:c:c1:3)c4:c:c:c:c:c2:4
mefenamic_acid	1	M01	Cc1:c:c:c:c(Nc2:c:c:c:c:c:2C(=O)O):c:1C
meloxicam	1	M01	CN1C(C(=O)Nc2:n:c:c(C):s:2)C(=O)c3:c:c:c:c:c:3S1(=O)=O
melperone	1	N05	CC1CCN(CCCC(=O)c2:c:c:c(F):c:c:2)CC1
meropenem	1	J01	CC(O)C1C2C(C)C(=C(N2C1=O)C(=O)O)SC3CNC(C3)C(=O)N(C)C
mesna	1	R05,V03	OS(=O)(=O)CCS
metamizole	1	N02	CN(CS(=O)(=O)O)C1=C(C)N(C)N(C1=O)c2:c:c:c:c:c:2
methazolamide	1	S01	CN1N=C(S/C/1=N/C(=O)C)S(=O)(=O)N
methotrexate	1	L01,L04	CN(Cc1:c:n:c2:n:c(N):n:c(N):c:2:n:1)c3:c:c:c(:c:c:3)C(=O)NC(CCC(=O)O)C(=O)O

Table A1.4.S1 (continued). Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
metildigoxin	1	C01	COC1C(O)CC(OC2C(O)CC(OC3C(O)CC(OC4CCC5(C)C(CCC6C5CC(O)C7(C)C(CCC67O)C8=CC(=O)OC8)C4)OC3C)OC2C)OC1C
metolazone	1	C03	CC1Nc2:c:c(Cl):c(:c:c:2C(=O)N1c3:c:c:c:c:c:3C)S(=O)(=O)N
metronidazole	1	A01,D06,G01,J01,P01	Cc1:n:c:c(N(=O)=O):n:1CCO
midecamycin	1	J01	CCC(=O)OC1CC(=O)OC(C)C\C=C/C=C/C(O)C(C)CC(CC=O)C(OC2OC(C)C(OC3CC(C)(O)C(OC(=O)CC)C(C)O3)C(C2O)N(C)C)C1OC
minocycline	1	A01,J01	CN(C)C1C2CC3Cc4:c(:c:c:c(O):c:4C(=O)C3C(=O)C2(O)C(=O)C(C(=O)N)C1=O)N(C)C
morniflumate	1	M01	FC(F)(F)c1:c:c:c:c(Nc2:n:c:c:c:c:2C(=O)OCCN3CCOCC3):c:1
moxifloxacin	1	J01,S01	COc1:c(N2CC3CCCNC3C2):c(F):c:c4C(=O)C(=CN(C5CC5)c:1:4)C(=O)O
nabumetone	1	M01	COc1:c:c:c2:c:c(CCC(=O)C):c:c:c:2:c:1
nevirapine	1	J05	Cc1:c:c:n:c2N(C3CC3)c4:n:c:c:c:c:4C(=O)Nc:1:2
nimesulide	1	M01,M02	CS(=O)(=O)Nc1:c:c:c(:c:c:1Oc2:c:c:c:c:c:2)N(=O)=O
norfloxacin	1	J01,S01	CCN1C=C(C(=O)O)C(=O)c2:c:c(F):c(:c:c1:2)N3CCNCC3
nystatin	1	A07,D01,G01	CC1OC(=O)CC(O)CC(O)CC(O)CCC(O)C(O)CC(=O)CC(O)C(C(O)CC(OC2OC(C)C(O)C(N)C2O)\C=C/C=C/C=C/C=C/CC\C=C\C=C\C(C)C(O)C1C)C(=O)O
oxaprozin	1	M01	OC(=O)CCc1:o:c(c2:c:c:c:c:c:2):c(:n:1)c3:c:c:c:c:c:3
oxazepam	1	N05	OC1N=C(c2:c:c:c:c:c:2)c3:c:c(Cl):c:c:c:3NC1=O
oxcarbazepine	1	N03	NC(=O)N1c2:c:c:c:c:c:2CC(=O)c3:c:c:c:c:c1:3
oxomemazine	1	R06	CC(CN(C)C)CN1c2:c:c:c:c:c:2S(=O)(=O)c3:c:c:c:c:c1:3
oxyphenbutazone	1	M01,M02,S01	CCCCC1C(=O)N(N(C1=O)c2:c:c:c(O):c:c:2)c3:c:c:c:c:c:3
oxytetracycline	1	D06,G01,J01,S01	CN(C)[C@@H]1[C@@H]2[C@H](O)[C@H]3C(C(=O)c4:c(O):c:c:c:c:4[C@]3(C)O)C(=O)[C@@]2(O)C(=O)C(C(=O)N)C1=O
pantoprazole	1	A02	COc1:c:c:n:c(CS(=O)c2:n:c3:c:c:c(OC(F)F):c:c:3:[nH]:2):c:1OC
paracetamol	1	N02	CC(=O)Nc1:c:c:c(O):c:c:1
pefloxacin	1	J01	CCN1C=C(C(=O)O)C(=O)c2:c:c(F):c(:c:c1:2)N3CCN(C)CC3
phenobarbital	1	N03	CCC1(C(=O)NC(=O)NC1=O)c2:c:c:c:c:c:2
phenoxymethylpenicillin	1	J01	CC1(C)SC2C(NC(=O)COc3:c:c:c:c:c:3)C(=O)N2C1C(=O)O
phenylbutazone	1	M01,M02	CCCCC1C(=O)N(N(C1=O)c2:c:c:c:c:c:2)c3:c:c:c:c:c:3
phenytoin	1	N03	O=C1NC(=O)C(N1)(c2:c:c:c:c:c:2)c3:c:c:c:c:c:3
phthalylsulfathiazole	1	A07	CCOC(=O)C1(CCN(CCc2:c:c:c(N):c:c:2)CC1)c3:c:c:c:c:c:3
piperacillin	1	J01	CCN1CCN(C(=O)NC(C(=O)NC2C3SC(C)(C)C(N3C2=O)C(=O)O)c4:c:c:c:c:c:4)C(=O)C1=O
pirenzepine	1	A02	CN1CCN(CC(=O)N2c3:c:c:c:c:c:3C(=O)Nc4:c:c:c:n:c2:4)CC1
piritramide	1	N02	NC(=O)C1(CCN(CCC(C#N)(c2:c:c:c:c:c:2)c3:c:c:c:c:c:3)CC1)N4CCCCC4
piroxicam	1	M01,M02,S01	CN1C(C(=O)Nc2:c:c:c:c:n:2)C(=O)c3:c:c:c:c:c:3S1(=O)=O
pivampicillin	1	J01	CC(C)(C)C(=O)OCOC(=O)C1N2C(SC1(C)C)C(NC(=O)C(N)c3:c:c:c:c:c:3)C2=O
prednisolone	1	A07,C05,D07,H02,R01,S01,S02,S03	CC12C[C@H](O)[C@H]3[C@H](CCC4=CC(=O)C=C[C@]34C)[C@@H]1CC[C@@]2(O)C(=O)CO
propylthiouracil	1	H03	CCCC1=NC(=S)NC(=O)C1
pyrazinamide	1	J04	NC(=O)c1:c:n:c:c:n:1
raltegravir	1	J05	CN1C(=O)C(=O)C(N=C1C(C)(C)NC(=O)c2:o:c(C):n:n:2)C(=O)NCc3:c:c:c(F):c:c:3
rifampicin	1	J04	COC1\C=C/OC2(C)Oc3:c(C):c(O):c4:c(O):c(NC(=O)\C(=C/C=C\C(C)C(O)C(C)C(O)C(C)C(OC(=O)C)C1C)\C):c(\C=N\N5CCN(C)CC5):c(O):c:4:c:3C2=O
roxithromycin	1	J01	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(O)C(C)O2)C(C)C(OC3OC(C)CC(C3O)N(C)C)C(C)(O)CC(C)\C(=N\OCOCCOC)\C(C)C(O)C1(C)O
sitagliptin	1	A10	NC(CC(=O)N1CCn2:c(C1):n:n:c:2C(F)(F)F)Cc3:c:c(F):c(F):c:c:3F
spiramycin	1	J01	COC1C(O)CC(=O)OC(C)C\C=C/C=C/C(OC2CCC(C(C)O2)N(C)C)C(C)CC(CC=O)C1OC3OC(C)C(OC4CC(C)(O)C(O)C(C)O4)C(C3C)N(C)C
spironolactone	1	C03	CC(=O)SC1CC2=CC(=O)CCC2(C)C3CCC4(C)[C@@H](CC[C@]45CCC(=O)O5)[C@@H]13
streptomycin	1	A07,J01	CNC1C(O)C(O)C(CO)OC1OC2C(OC3C(O)C(O)C(N=C(N)N)C(O)C3N=C(N)N)OC(C)C2(O)C=O
sulfacetamide	1	S01	CC(=O)NS(=O)(=O)c1:c:c:c(N):c:c:1
sulfadiazine	1	J01	Nc1:c:c:c(:c:c:1)S(=O)(=O)Nc2:n:c:c:c:n:2
sulfadimethoxine	1	J01	COc1:c:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):n:c(OC):n:1
sulfadimidine	1	J01	Cc1:c:c(C):n:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):n:1
sulfadoxine	1	J01	COc1:n:c:n:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):c:1OC
sulfafurazole	1	J01,S01	Cc1:n:o:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):c:1C
sulfaguanidine	1	A07	NC(=N)NS(=O)(=O)c1:c:c:c(N):c:c:1
sulfalene	1	J01	COc1:n:c:c:n:c:1NS(=O)(=O)c2:c:c:c(N):c:c:2
sulfamethoxazole	1	J01	Cc1:o:n:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):c:1
sulfamethoxypyridazine	1	J01	COc1:c:c:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):n:n:1
sulfametoxydiazine	1	J01	COc1:c:n:c(NS(=O)(=O)c2:c:c:c(N):c:c:2):n:c:1
sulfasalazine	1	A07	OC(=O)c1:c:c(:c:c:c:1O)N=Nc2:c:c:c(:c:c:2)S(=O)(=O)Nc3:c:c:c:c:n:3
sulindac	1	M01	CC1=C(CC(=O)O)c2:c:c(F):c:c:c:2/C/1=C/c3:c:c:c(:c:c:3)S(=O)C
sultamicillin	1	J01	CC1(C)SC2C(NC(=O)C(N)c3:c:c:c:c:c:3)C(=O)N2C1C(=O)OCOC(=O)C4N5C(CC5=O)S(=O)(=O)C4(C)C
temozolomide	1	L01	CN1N=Nc2:c(:n:c:n:2C1=O)C(=O)N
tenoxicam	1	M01	CN1C(C(=O)Nc2:c:c:c:c:n:2)C(=O)c3:s:c:c:c:3S1(=O)=O
terbinafine	1	D01	CN(C\C=C\C#CC(C)(C)C)Cc1:c:c:c:c2:c:c:c:c:c:1:2
tetracycline	1	A01,D06,J01,S01,S02,S03	CN(C)C1C2CC3C(C(=O)c4:c(O):c:c:c:c:4C3(C)O)C(=O)C2(O)C(=O)C(C(=O)N)C1=O

Table A1.4.S1 (continued). Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
tetrazepam	1	M03	CN1C(=O)CN=C(C2=CCCCC2)c3:c:c(Cl):c:c:c1:3
thioacetazone	1	J04	CC(=O)Nc1:c:c:c(\C=N\NC(=S)N):c:c:1
tiabendazole	1	D01,P02	c1:c:c:c2:[nH]:c(:n:c:2:c:1)c3:c:s:c:n:3
tiaprofenic_acid	1	M01	CC(C(=O)O)c1:c:c:c(:s:1)C(=O)c2:c:c:c:c:c:2
tocainide	1	C01	CC(N)C(=O)Nc1:c(C):c:c:c:c:1C
torasemide	1	C03	CC(C)NC(=O)NS(=O)(=O)c1:c:n:c:c:c:1Nc2:c:c:c:c(C):c:2
trihexyphenidyl	1	N04	OC(CCN1CCCCC1)(C2CCCCC2)c3:c:c:c:c:c:3
trimethoprim	1	J01	COc1:c:c(Cc2:c:n:c(N):n:c:2N):c:c(OC):c:1OC
valdecoxib	1	M01	Cc1:o:n:c(c2:c:c:c:c:c:2):c:1c3:c:c:c(:c:c:3)S(=O)(=O)N
vancomycin	1	A07,J01	CNC(CC(C)C)C(=O)NC1C(O)c2:c:c:c(Oc3:c:c4:c:c(Oc5:c:c:c(:c:c:5Cl)C(O)C6NC(=O)C(NC(=O)C4NC(=O)C(CC(=O)N)NC1=O)c7:c:c:c(O):c(:c:7)c8:c(O):c:c(O):c:c:8C(NC6=O)C(=O)O):c:3OC9OC(CO)C(O)C(O)C9OC%10CC(C)(N)C(O)C(C)O%10):c(Cl):c:2
verapamil	1	C08	COc1:c:c:c(CCN(C)CCCC(C#N)(C(C)C)c2:c:c:c(OC):c(OC):c:2):c:c:1OC
voriconazole	1	J02	CC(c1:n:c:n:c:c:1F)C(O)(Cn2:c:n:c:n:2)c3:c:c:c(F):c:c:3F
xylometazoline	1	R01,S01	Cc1:c:c(:c:c(C):c:1CC2=NCCN2)C(C)(C)C
zonisamide	1	N03	NS(=O)(=O)Cc1:n:o:c2:c:c:c:c:c:1:2
acebutolol	0	C07	CCCC(=O)Nc1:c:c:c(OCC(O)CNC(C)C):c(:c:1)C(=O)C
acenocoumarol	0	B01	CC(=O)CC(C1C(=O)Oc2:c:c:c:c:c:2C1=O)c3:c:c:c(:c:c:3)N(=O)=O
adenosine	0	C01	Nc1:n:c:n:c2:c:1:n:c:n:2C3OC(CO)C(O)C3O
agomelatine	0	N06	COc1:c:c:c2:c:c:c:c(CCNC(=O)C):c:2:c:1
alfuzosin	0	G04	COc1:c:c2:n:c(:n:c(N):c:2:c:c:1OC)N(C)CCCNC(=O)C3CCCO3
alosetron	0	A03	Cc1:[nH]:c:n:c:1CN2CCc3:c(C2=O):c4:c:c:c:c:c:4:n:3C
alprostadil	0	C01,G04	CCCCCC(O)\C=C\C1C(O)CC(=O)C1CCCCCCC(=O)O
amfepramone	0	A08	CCN(CC)C(C)C(=O)c1:c:c:c:c:c:1
atracurium	0	M03	COc1:c:c:c(CC2c3:c:c(OC):c(OC):c:c:3CCN2(C)CCC(=O)OCCCCCOC(=O)CCN4(C)CCc5:c:c(OC):c(OC):c:c:5C4Cc6:c:c:c(OC):c(OC):c:6):c:c:1OC
azacitidine	0	L01	NC1=NC(=O)N(C=N1)C2OC(CO)C(O)C2O
azelastine	0	R01,R06,S01	CN1CCCC(CC1)N2N=C(Cc3:c:c:c(Cl):c:c:3)c4:c:c:c:c:c:4C2=O
beclometasone	0	A07,D07,R01,R03	CC1CC2C3CCC4=CC(=O)C=CC4(C)C3(Cl)C(O)CC2(C)C1(O)C(=O)CO
benzatropine	0	N04	CN1C2CCC1CC(C2)OC(c3:c:c:c:c:c:3)c4:c:c:c:c:c:4
benzoyl_peroxide	0	D10	O=C(OOC(=O)c1:c:c:c:c:c:1)c2:c:c:c:c:c:2
bleomycin	0	L01	CC(O)C(NC(=O)C(C)C(O)C(C)NC(=O)C(NC(=O)c1:n:c(:n:c(N):c:1C)C(CC(=O)N)NCC(N)C(=O)N)C(OC2OC(CO)C(O)C(O)C2OC3OC(CO)C(O)C(OC(=O)N)C3O)c4:c:[nH]:c:n:4)C(=O)NCCc5:n:c(:c:s:5)c6:n:c(:c:s:6)C(=O)NCCCS(C)C
bosentan	0	C02	COc1:c:c:c:c:c:1Oc2:c(NS(=O)(=O)c3:c:c:c(:c:c:3)C(C)(C)C):n:c(:n:c:2OCCO)c4:n:c:c:c:n:4
brimonidine	0	S01	Brc1:c(NC2=NCCN2):c:c:c3:n:c:c:n:c:1:3
bromfenac	0	S01	Nc1:c(CC(=O)O):c:c:c:c:1C(=O)c2:c:c:c(Br):c:c:2
bromocriptine	0	G02,N04	CC(C)CC1N2C(=O)C(NC(=O)C3CN(C)C4Cc5:c(Br):[nH]:c6:c:c:c:c(C4=C3):c:5:6)(OC2(O)C7CCCN7C1=O)C(C)C
bupivacaine	0	N01	CCCCN1CCCCC1C(=O)Nc2:c(C):c:c:c:c:2C
busulfan	0	L01	CS(=O)(=O)OCCCCOS(=O)(=O)C
butorphanol	0	N02	Oc1:c:c:c2CC3N(CC4CCC4)CCC5(CCCCC35O)c:2:c:1
cabergoline	0	G02,N04	CCNC(=O)N(CCCN(C)C)C(=O)C1CC2C(Cc3:c:[nH]:c4:c:c:c:c2:c:3:4)N(CC=C)C1
carmustine	0	L01	ClCCNC(=O)N(CCCl)N=O
cetirizine	0	R06	OC(=O)COCCN1CCN(CC1)C(c2:c:c:c:c:c:2)c3:c:c:c(Cl):c:c:3
cilazapril	0	C09	CCOC(=O)C(CCc1:c:c:c:c:c:1)NC2CCCN3CCCC(N3C2=O)C(=O)O
cinacalcet	0	H05	CC(NCCCc1:c:c:c:c(:c:1)C(F)(F)F)c2:c:c:c:c3:c:c:c:c:c:2:3
cisapride	0	A03	COC1CN(CCCOc2:c:c:c(F):c:c:2)CCC1NC(=O)c3:c:c(Cl):c(N):c:c:3OC
clobetasol	0	D07	CCC(=O)OC1(C(C)CC2C3CCC4=CC(=O)C=CC4(C)C3(F)C(O)CC12C)C(=O)CCl
clodronic_acid	0	M05	OP(=O)(O)C(Cl)(Cl)P(=O)(O)O
clonazepam	0	N03	Clc1:c:c:c:c:c:1C2=NCC(=O)Nc3:c:c:c(:c:c2:3)N(=O)=O
cocaine	0	N01,R02,S01,S02	COC(=O)C1C(CC2CCC1N2C)OC(=O)c3:c:c:c:c:c:3
cromoglicic_acid	0	A07,D11,R01,R03,S01	OC(COc1:c:c:c:c2OC(=CC(=O)c:1:2)C(=O)O)COc3:c:c:c:c4OC(=CC(=O)c:3:4)C(=O)O
cyamemazine	0	N05	CC(CN(C)C)CN1c2:c:c:c:c:c:2Sc3:c:c:c(:c:c1:3)C#N
cyclizine	0	R06	CN1CCN(CC1)C(c2:c:c:c:c:c:2)c3:c:c:c:c:c:3
cycloserine	0	J04	NC1CONC1=O
cyproterone	0	G03	CC(=O)C1(O)CCC2C3C=C(Cl)C4=CC(=O)C5CC5C4(C)C3CCC12C
dacarbazine	0	L01	CN(C)N=Nc1:n:c:[nH]:c:1C(=O)N
daunorubicin	0	L01	COc1:c:c:c:c2C(=O)c3:c(O):c4CC(O)(CC(OC5CC(N)C(O)C(C)O5)c:4:c(O):c:3C(=O)c:1:2)C(=O)C
desogestrel	0	G03	CCC12CC(=C)C3C(CCC4=CCCCC34)C1CCC2(O)C#C
dexamfetamine	0	N06	CC(N)Cc1:c:c:c:c:c:1
dextran	0	B05	OCC1OC(OCC2OC(OCC(O)C(O)C(O)C(O)C=O)C(O)C(O)C2O)C(O)C(O)C1O
dicycloverine	0	A03	CCN(CC)CCOC(=O)C1(CCCCC1)C2CCCCC2
diethylstilbestrol	0	G03,L02	CC\C(=C(\CC)/c1:c:c:c(O):c:c:1)\c2:c:c:c(O):c:c:2
digoxin	0	C01	CC1OC(CC(O)C1O)OC2C(O)CC(OC3C(O)CC(OC4CCC5(C)C(CCC6C5CC(O)C7(C)C(CCC67O)C8=CC(=O)OC8)C4)OC3C)OC2C

Table A1.4.S1 (continued). Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
dinoprostone	0	G02	CCCCCC(O)\C=C\C1C(O)CC(=O)C1C\C=C\CCCC(=O)O
diosmin	0	C05	COc1:c:c:c(:c:c:1O)C2=CC(=O)c3:c(O):c:c(OC4OC(COC5OC(C)C(O)C(O)C5O)C(O)C(O)C4O):c:c:3O2
dipivefrine	0	S01	CNCC(O)c1:c:c:c(OC(=O)C(C)(C)C):c(OC(=O)C(C)(C)C):c:1
dipyridamole	0	B01	OCCN(CCO)c1:n:c(N2CCCCC2):c3:n:c(:n:c(N4CCCCC4):c:3:n:1)N(CCO)CCO
disopyramide	0	C01	CC(C)N(CCC(C(=O)N)(c1:c:c:c:c:c:1)c2:c:c:c:c:n:2)C(C)C
docosanol	0	D06	CCCCCCCCCCCCCCCCCCCCCCO
dofetilide	0	C01	CN(CCOc1:c:c:c(NS(=O)(=O)C):c:c:1)CCc2:c:c:c(NS(=O)(=O)C):c:c:2
donepezil	0	N06	COc1:c:c2CC(CC3CCN(Cc4:c:c:c:c:c:4)CC3)C(=O)c:2:c:c:1OC
droperidol	0	N01,N05	Fc1:c:c:c(:c:c:1)C(=O)CCCN2CCC(=CC2)N3C(=O)Nc4:c:c:c:c:c3:4
eletriptan	0	N02	CN1CCCC1Cc2:c:[nH]:c3:c:c:c(CCS(=O)(=O)c4:c:c:c:c:c:4):c:c:2:3
eltrombopag	0	B02	Cc1:c:c:c(:c:c:1C)n2:n:c(C):c(N=Nc3:c:c:c:c(:c:3O)c4:c:c:c:c(:c:4)C(=O)O):c:2O
encainide	0	C01	COc1:c:c:c(:c:c:1)C(=O)Nc2:c:c:c:c:c:2CCC3CCCCN3C
entacapone	0	N04	CCN(CC)C(=O)\C(=C/c1:c:c(O):c(O):c(:c:1)N(=O)=O)\C#N
entecavir	0	J05	NC1=Nc2:c(:n:c:n:2C3CC(O)C(CO)C3=C)C(=O)N1
epinephrine	0	A01,B02,C01,R01,R03,S01	CNCC(O)c1:c:c:c(O):c(O):c:1
epirubicin	0	L01	COc1:c:c:c:c2C(=O)c3:c(O):c4CC(O)(CC(OC5CC(N)C(O)C(C)O5)c:4:c(O):c:3C(=O)c:1:2)C(=O)CO
eprosartan	0	C09	CCCCc1:n:c:c(\C=C(\Cc2:c:c:c:s:2)/C(=O)O):n:1Cc3:c:c:c(:c:c:3)C(=O)O
eptifibatide	0	B01	NC(=N)NCCCCC1NC(=O)CCSSCC(NC(=O)C2CCCN2C(=O)C(Cc3:c:[nH]:c4:c:c:c:c:c:3:4)NC(=O)C(CC(=O)O)NC(=O)CNC1=O)C(=O)N
eszopiclone	0	N05	CN1CCN(CC1)C(=O)OC2N(C(=O)c3:n:c:c:n:c2:3)c4:c:c:c(Cl):c:n:4
ethanol	0	D08,V03	CCO
ethinylestradiol	0	G03,L02	CC12CCC3C(CCc4:c:c(O):c:c:c3:4)C1CCC2(O)C#C
everolimus	0	L01,L04	COC1CC(CC(C)C2CC(=O)\C(=C\C(C)C(O)C(OC)C(=O)C(C)CC(C)\C=C\C=C\C=C(\C)/C(CC3CCC(C)C(O)(O3)C(=O)C(=O)N4CCCCC4C(=O)O2)OC)\C)CCC1OCCO
fampridine	0	N07	Nc1:c:c:n:c:c:1
fenfluramine	0	A08	CCNC(C)Cc1:c:c:c:c(:c:1)C(F)(F)F
gadobutrol	0	V08	OCC(O)C(CO)N1CCN(CC(=O)O)CCN(CC(=O)O)CCN(CC(=O)O)CC1
gadoteridol	0	V08	CC(O)CN1CCN(CC(=O)O)CCN(CC(=O)O)CCN(CC(=O)O)CC1
galantamine	0	N06	COc1:c:c:c2CN(C)CCC34C=CC(O)CC3Oc:1:c:24
glafenine	0	N02	OCC(O)COC(=O)c1:c:c:c:c:c:1Nc2:c:c:n:c3:c:c(Cl):c:c:c:2:3
glyceryl_trinitrate	0	C01,C05	O=N(=O)OCC(CON(=O)=O)ON(=O)=O
haloperidol	0	N05	OC1(CCN(CCCC(=O)c2:c:c:c(F):c:c:2)CC1)c3:c:c:c(Cl):c:c:3
hydrocodone	0	R05	COc1:c:c:c2CC3C4CCC(=O)C5Oc:1:c:2C45CCN3C
idarubicin	0	L01	CC1OC(CC(N)C1O)OC2CC(O)(Cc3:c(O):c4C(=O)c5:c:c:c:c:c:5C(=O)c:4:c(O):c2:3)C(=O)C
ifosfamide	0	L01	ClCCNP1(=O)OCCCN1CCCl
iobitridol	0	V08	CN(CC(O)CO)C(=O)c1:c(I):c(NC(=O)C(CO)CO):c(I):c(C(=O)N(C)CC(O)CO):c:1I
iomeprol	0	V08	CN(C(=O)CO)c1:c(I):c(C(=O)NCC(O)CO):c(I):c(C(=O)NCC(O)CO):c:1I
iotalamic_acid	0	V08	CNC(=O)c1:c(I):c(NC(=O)C):c(I):c(C(=O)O):c:1I
ioversol	0	V08	OCCN(C(=O)CO)c1:c(I):c(C(=O)NCC(O)CO):c(I):c(C(=O)NCC(O)CO):c:1I
irinotecan	0	L01	CCc1:c2CN3C(=O)C4=C(C=C3c:2:n:c5:c:c:c(OC(=O)N6CCC(CC6)N7CCCCC7):c:c:1:5)C(O)(CC)C(=O)OC4
ixabepilone	0	L01	CC1CCCC2(C)OC2CC(NC(=O)CC(O)C(C)(C)C(=O)C(C)C1O)\C(=C\c3:c:s:c(C):n:3)\C
ketotifen	0	R06,S01	CN1CCC(=C2c3:c:c:c:c:c:3CC(=O)c4:s:c:c:c2:4)CC1
lacidipine	0	C08	CCCCc1:n:c(Cl):c(CO):n:1Cc2:c:c:c(:c:c:2)c3:c:c:c:c:c:3c4:n:n:n:[nH]:4
latamoxef	0	J01	COC1(NC(=O)C(C(=O)O)c2:c:c:c(O):c:c:2)C3OCC(=C(N3C1=O)C(=O)O)CSc4:n:n:n:n:4C
levonorgestrel	0	G03	CCC12CCC3C(CCC4=CC(=O)CCC34)C1CCC2(O)C#C
loxapine	0	N05	CN1CCN(CC1)C2=Nc3:c:c:c:c:c:3Oc4:c:c:c(Cl):c:c2:4
melphalan	0	L01	NC(Cc1:c:c:c(:c:c:1)N(CCCl)CCCl)C(=O)O
metamfetamine	0	N06	CNC(C)Cc1:c:c:c:c:c:1
methadone	0	N07	CCC(=O)C(CC(C)N(C)C)(c1:c:c:c:c:c:1)c2:c:c:c:c:c:2
methysergide	0	N02	CCC(CO)NC(=O)C1CN(C)C2Cc3:c:n(C):c4:c:c:c:c(C2=C1):c:3:4
metrizamide	0	V08	CN(C(=O)C)c1:c(I):c(NC(=O)C):c(I):c(C(=O)NC2C(O)OC(CO)C(O)C2O):c:1I
mibefradil	0	C08	COCC(=O)OC1(CCN(C)CCCc2:n:c3:c:c:c:c:c:3:[nH]:2)CCc4:c:c(F):c:c:c:4C1C(C)C
mifepristone	0	G03	CC#CC1(O)CCC2C3CCC4=CC(=O)CCC4=C3C(CC12C)c5:c:c:c(:c:c:5)N(C)C
misoprostol	0	A02,G02	CCCCC(C)(O)C\C=C\C1C(O)CC(=O)C1CCCCCCC(=O)OC
mitomycin	0	L01	COC12C3NC3CN1C4=C(C2COC(=O)N)C(=O)C(=N)C(C)C4=O
mitoxantrone	0	L01	OCCNCCNc1:c:c:c(NCCNCCO):c2C(=O)c3:c(O):c:c:c(O):c:3C(=O)c:1:2
moclobemide	0	N06	Clc1:c:c:c(:c:c:1)C(=O)NCCN2CCOCC2
mometasone	0	D07,R01,R03	CC1CC2C3CCC4=CC(=O)C=CC4(C)C3(Cl)C(O)CC2(C)C1(O)C(=O)CCl
nafarelin	0	H01	CC(C)CC(\N=C(\O)/C(Cc1:c:c:c2:c:c:c:c:c:2:c:1)\N=C(\O)/C(Cc3:c:c:c(O):c:c:3)\N=C(\O)/C(CO)\N=C(\O)/C(Cc4:c:[nH]:c5:c:c:c:c:c:4:5)\N=C(\O)/C(Cc6:c:[nH]:c:n:6)\N=C(\O)/C7CCC(=N7)O)\C(=N\C(CCCNC(=N)N)C(=O)N8CCCC8\C(=N\CC(=N)O)\O)\O
nalbuphine	0	N02	OC1CCC2(O)C3Cc4:c:c:c(O):c5OC1C2(CCN3CC6CCC6)c:4:5
naltrexone	0	N07	Oc1:c:c:c2CC3N(CC4CC4)CCC56C(Oc:1:c:25)C(=O)CCC36O
nilotinib	0	L01	Cc1:c:n(:c:n:1)c2:c:c(NC(=O)c3:c:c:c(C):c(Nc4:n:c:c:c(:n:4)c5:c:c:c:n:c:5):c:3):c:c(:c:2)C(F)(F)F

Table A1.4.S1 (continued). Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
nizatidine	0	A02	C\N=C(/CN(=O)=O)\NCCSCc1:c:s:c(CN(C)C):n:1
norethisterone	0	G03	CC12CCC3C(CCC4=CC(=O)CCC34)C1CCC2(O)C#C
nortriptyline	0	N06	CNCCC=C1c2:c:c:c:c:c:2CCc3:c:c:c:c:c1:3
octreotide	0	H01	CC(O)C(CO)NC(=O)C1CSSCC(NC(=O)C(N)Cc2:c:c:c:c:c:2)C(=O)NC(Cc3:c:c:c:c:c:3)C(=O)NC(Cc4:c:[nH]:c5:c:c:c:c:c:4:5)C(=O)NC(CCCCN)C(=O)NC(C(C)O)C(=O)N1
ondansetron	0	A04	Cc1:n:c:c:n:1CC2CCc3:c(C2=O):c4:c:c:c:c:c:4:n:3C
orciprenaline	0	R03	CC(C)NCC(O)c1:c:c(O):c:c(O):c:1
orlistat	0	A08	CCCCCCCCCCCC(CC1OC(=O)C1CCCCCC)OC(=O)C(CC(C)C)NC=O
oxybutynin	0	G04	CCN(CC)CC#CCOC(=O)C(O)(C1CCCCC1)c2:c:c:c:c:c:2
oxytocin	0	H01	CCC(C)C1NC(=O)C(Cc2:c:c:c(O):c:c:2)NC(=O)C(N)CSSCC(NC(=O)C(CC(=O)N)NC(=O)C(CCC(=O)N)NC1=O)C(=O)N3CCCC3C(=O)NC(CC(C)C)C(=O)NCC(=O)N
pergolide	0	N04	CCCN1CC(CSC)CC2C1Cc3:c:[nH]:c4:c:c:c:c2:c:3:4
perindopril	0	C09	CCCC(NC(C)C(=O)N1C2CCCCC2CC1C(=O)O)C(=O)OCC
permethrin	0	P03	CC1(C)C(C=C(Cl)Cl)C1C(=O)OCc2:c:c:c:c(Oc3:c:c:c:c:c:3):c:2
pethidine	0	N02	CCOC(=O)C1(CCN(C)CC1)c2:c:c:c:c:c:2
phenylephrine	0	C01,R01,S01	CNCC(O)c1:c:c:c:c(O):c:1
phenylpropanolamine	0	R01	CC(N)C(O)c1:c:c:c:c:c:1
pimecrolimus	0	D11	CC\C\1=C/C(C)CC(C)CC(OC)C2OC(O)(C(C)CC2OC)C(=O)C(=O)N3CCCCC3C(=O)OC(C(C)C(O)CC1=O)\C(=C\C4CCC(Cl)C(C4)OC)\C
pimozide	0	N05	Fc1:c:c:c(:c:c:1)C(CCCN2CCC(CC2)N3C(=O)Nc4:c:c:c:c:c3:4)c5:c:c:c(F):c:c:5
pindolol	0	C07	CC(C)NCC(O)COc1:c:c:c:c2:[nH]:c:c:c:1:2
pizotifen	0	N02	CN1CCC(=C2c3:c:c:c:c:c:3CCc4:s:c:c:c2:4)CC1
prasugrel	0	B01	CC(=O)Oc1:c:c2CN(CCc:2:s:1)C(C(=O)C3CC3)c4:c:c:c:c:c:4F
prazosin	0	C02	COc1:c:c2:n:c(:n:c(N):c:2:c:c:1OC)N3CCN(CC3)C(=O)c4:o:c:c:c:4
prilocaine	0	N01	CCCNC(C)C(=O)Nc1:c:c:c:c:c:1C
probucol	0	C10	CC(C)(C)c1:c:c(SC(C)(C)Sc2:c:c(:c(O):c(:c:2)C(C)(C)C)C(C)(C)C):c:c(:c:1O)C(C)(C)C
progesterone	0	G03	CC(=O)C1CCC2C3CCC4=CC(=O)CCC4(C)C3CCC12C
propafenone	0	C01	CCCNCC(O)COc1:c:c:c:c:c:1C(=O)CCc2:c:c:c:c:c:2
quinapril	0	C09	CCOC(=O)C(CCc1:c:c:c:c:c:1)NC(C)C(=O)N2Cc3:c:c:c:c:c:3CC2C(=O)O
rasagiline	0	N04	C#CCNC1CCc2:c:c:c:c:c1:2
retapamulin	0	D06	CC1CCC23CCC(=O)C2C1(C)C(CC(C)(C=C)C(O)C3C)OC(=O)CSC4CC5CCC(C4)N5C
rimonabant	0	A08	Cc1:c(:n:n(c2:c:c:c(Cl):c:c:2Cl):c:1c3:c:c:c(Cl):c:c:3)C(=O)NN4CCCCC4
ritodrine	0	G02	CC(NCCc1:c:c:c(O):c:c:1)C(O)c2:c:c:c(O):c:c:2
rizatriptan	0	N02	CN(C)CCc1:c:[nH]:c2:c:c:c(Cn3:c:n:c:n:3):c:c:1:2
ropinirole	0	N04	CCCN(CCC)CCc1:c:c:c:c2NC(=O)Cc:1:2
ropivacaine	0	N01	CCCN1CCCCC1C(=O)Nc2:c(C):c:c:c:c:2C
rotigotine	0	N04	CCCN(CCc1:c:c:c:s:1)C2CCc3:c(O):c:c:c:c:3C2
selegiline	0	N04	CC(Cc1:c:c:c:c:c:1)N(C)CC#C
sevoflurane	0	N01	FCOC(C(F)(F)F)C(F)(F)F
sotalol	0	C07	CC(C)NCC(O)c1:c:c:c(NS(=O)(=O)C):c:c:1
suxamethonium	0	M03	CN(C)(C)CCOC(=O)CCC(=O)OCCN(C)(C)C
temafloxacin	0	J01	CC1CN(CCN1)c2:c:c3N(C=C(C(=O)O)C(=O)c:3:c:c:2F)c4:c:c:c(F):c:c:4F
terazosin	0	G04	COc1:c:c2:n:c(:n:c(N):c:2:c:c:1OC)N3CCN(CC3)C(=O)C4CCCO4
terconazole	0	G01	CC(C)N1CCN(CC1)c2:c:c:c(OCC3COC(Cn4:c:n:c:n:4)(O3)c5:c:c:c(Cl):c:c:5Cl):c:c:2
thiethylperazine	0	R06	CCSc1:c:c:c2Sc3:c:c:c:c:c:3N(CCCN4CCN(C)CC4)c:2:c:1
thiopental	0	N01,N05	CCCC(C)C1(CC)C(=O)NC(=S)NC1=O
tiagabine	0	N03	Cc1:c:c:s:c:1C(=CCCN2CCCC(C2)C(=O)O)c3:s:c:c:c:3C
tibolone	0	G03	CC1CC2=C(CCC(=O)C2)C3CCC4(C)[C@@H](CC[C@@]4(O)C#C)[C@@H]13
tioconazole	0	D01,G01	Clc1:c:c:c(C(Cn2:c:c:n:c:2)OCc3:c:c:s:c:3Cl):c(Cl):c:1
tirofiban	0	B01	CCCCS(=O)(=O)NC(Cc1:c:c:c(OCCCCC2CCNCC2):c:c:1)C(=O)O
topotecan	0	L01	CCC1(O)C(=O)OCC2=C1C=C3N(Cc4:c:c5:c(CN(C)C):c(O):c:c:c:5:n:c3:4)C2=O
tranylcypromine	0	N06	NC1CC1c2:c:c:c:c:c:2
travoprost	0	S01	CC(C)OC(=O)CCC\C=C\CC1C(O)CC(O)C1\C=C\C(O)COc2:c:c:c:c(:c:2)C(F)(F)F
treprostinil	0	B01	CCCCCC(O)CCC1C(O)CC2Cc3:c(CC12):c:c:c:c:3OCC(=O)O
tretinoin	0	D10,L01	C\C(=C/C=C/C(=C/C(=O)O)/C)\C=C\C1=C(C)CCCC1(C)C
triamcinolone	0	A01,C05,D07,H02,R01,R03,S01	CC12CC(O)C3(F)C(CCC4=CC(=O)C=CC34C)C1CC(O)C2(O)C(=O)CO
triazolam	0	N05	Cc1:n:n:c2CN=C(c3:c:c:c:c:c:3Cl)c4:c:c(Cl):c:c:c:4n:1:2
valganciclovir	0	J05	CC(C)C(N)C(=O)OCC(CO)OCn1:c:n:c2:c(O):n:c(N):n:c:1:2
vardenafil	0	G04	CCCc1:n:c(C):c2:c(O):n:c(:n:n:1:2)c3:c:c(:c:c:c:3OCC)S(=O)(=O)N4CCN(CC)CC4
vecuronium	0	M03	CC(=O)OC1CC2CCC3C(CCC4(C)C3CC(C4OC(=O)C)N5(C)CCCCC5)C2(C)CC1N6CCCCC6
verteporfin	0	S01	COC(=O)CCc1:c(C):c2:c:c3:n:c(:c:c4:n:c(:c:c5:n:c(:c:c:1:n:2):c(CCC(=O)O):c:5C):c(C=C):c:4C)C6=CC=C(C(C(=O)OC)C36C)C(=O)OC

Table A1.4.S1 (continued). Drugs used for modeling

Name	SJS activity	ATC codes	Canonical Smiles
vinblastine	0	L01	CCC1(O)CC2CN(CCc3:c(:[nH]:c4:c:c:c:c:c:3:4)C(C2)(C(=O)OC)c5:c:c6:c(:c:c:5OC)N(C)C7C(O)(C(OC(=O)C)C8(CC)C=CCN9CCC67C89)C(=O)OC)C1
vincristine	0	L01	CCC1(O)CC2CN(CCc3:c(:[nH]:c4:c:c:c:c:c:3:4)C(C2)(C(=O)OC)c5:c:c6:c(:c:c:5OC)N(C=O)C7C(O)(C(OC(=O)C)C8(CC)C=CCN9CCC67C89)C(=O)OC)C1
vinorelbine	0	L01	CCC1=CC2CN(C1)Cc3:c(:[nH]:c4:c:c:c:c:c:3:4)C(C2)(C(=O)OC)c5:c:c6:c(:c:c:5OC)N(C)C7C(O)(C(OC(=O)C)C8(CC)C=CCN9CCC67C89)C(=O)OC
zafirlukast	0	R03	COc1:c:c(:c:c:c:1Cc2:c:n(C):c3:c:c:c(NC(=O)OC4CCCC4):c:c:2:3)C(=O)NS(=O)(=O)c5:c:c:c:c:c:5C
zaleplon	0	N05	CCN(C(=O)C)c1:c:c:c:c(:c:1)c2:c:c:n:c3:c(:c:n:n:2:3)C#N
zolpidem	0	N05	CN(C)C(=O)Cc1:c(:n:c2:c:c:c(C):c:n:1:2)c3:c:c:c(C):c:c:3
zuclopenthixol	0	N05	OCCN1CCN(CC\C=C\2/c3:c:c:c:c:c:3Sc4:c:c:c(Cl):c:c2:4)CC1

Table A1.4.S2. Chemical descriptors used for QSAR modeling

(only the first 10 rows of over 1000 rows are shown; table available upon request)

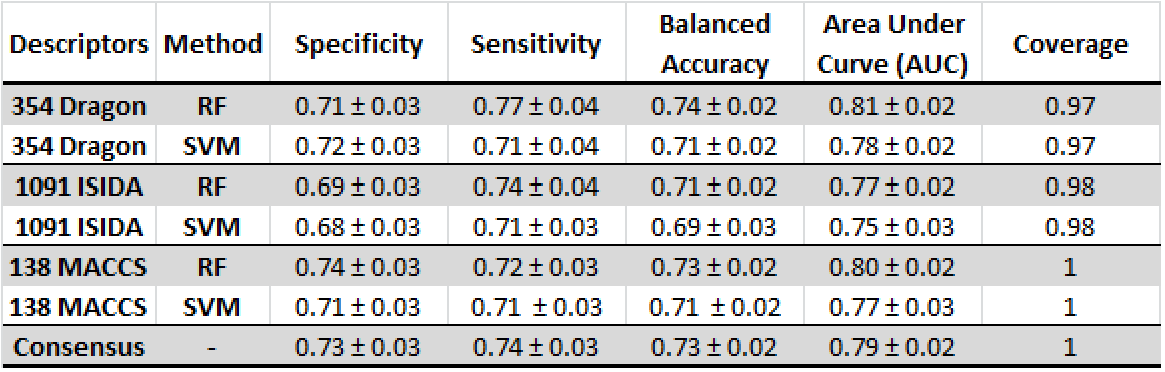
Dragon descriptors	ISIDA fragments[a]	MACCS fingerprints
MW	H-C-C-N=C	8_QAAA@1
AMW	C-C-C-C-C	11_4MRING
Ss	H-C-C-N-C-C	17_CTC
Mv	C-C*C*C*C-O	19_7MRING
Me	C-C*C*C*C-N	22_3MRING
Mp	C*C*C-C-C-N	23_NC(O)O
Ms	H-C-C-N-C-H	24_N-O
nBM	C*C*C-C-C-O	25_NC(N)N
SCBO	H-C=C-N-C=O	26_C$=C($A)$A

[a] types of ISIDA bonds

-	single bond
=	double bond
*	aromatic bond

For explanation of descriptors, refer to:
Dragon: http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf
ISIDA: http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/Nomenclature_of_ISIDA_fragments_2011.pdf
MACCS: http://www.mayachemtools.org/docs/scripts/html/MACCSKeysFingerprints.html

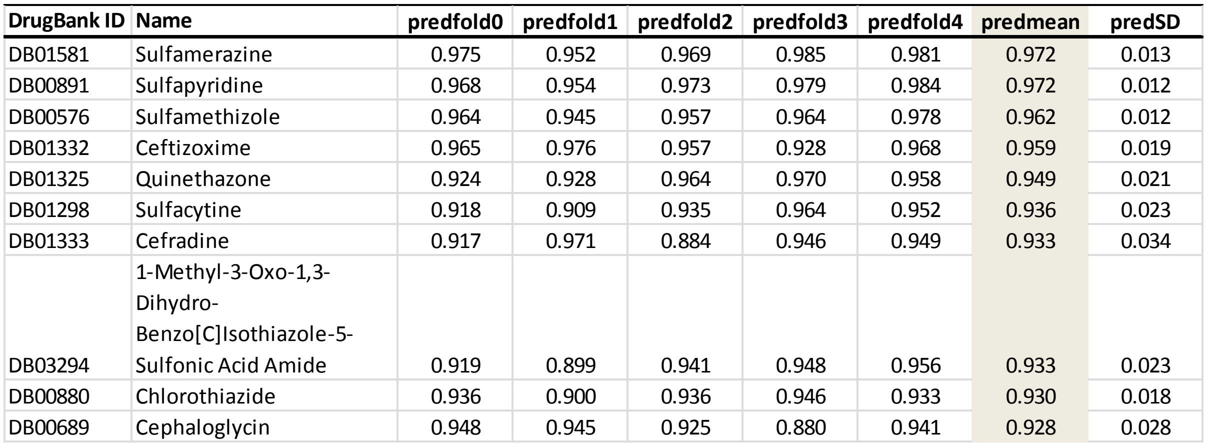
Table A1.4.S3. Performance of QSAR models



Descriptors	Method	Specificity	Sensitivity	Balanced Accuracy	Area Under Curve (AUC)	Coverage
354 Dragon	RF	0.71 ± 0.03	0.77 ± 0.04	0.74 ± 0.02	0.81 ± 0.02	0.97
354 Dragon	SVM	0.72 ± 0.03	0.71 ± 0.04	0.71 ± 0.02	0.78 ± 0.02	0.97
1091 ISIDA	RF	0.69 ± 0.03	0.74 ± 0.04	0.71 ± 0.02	0.77 ± 0.02	0.98
1091 ISIDA	SVM	0.68 ± 0.03	0.71 ± 0.03	0.69 ± 0.03	0.75 ± 0.03	0.98
138 MACCS	RF	0.74 ± 0.03	0.72 ± 0.03	0.73 ± 0.02	0.80 ± 0.02	1
138 MACCS	SVM	0.71 ± 0.03	0.71 ± 0.03	0.71 ± 0.02	0.77 ± 0.03	1
Consensus	-	0.73 ± 0.03	0.74 ± 0.03	0.73 ± 0.02	0.79 ± 0.02	1

Table A1.4.S4. Predictions of SJS activity of drugs in DrugBank

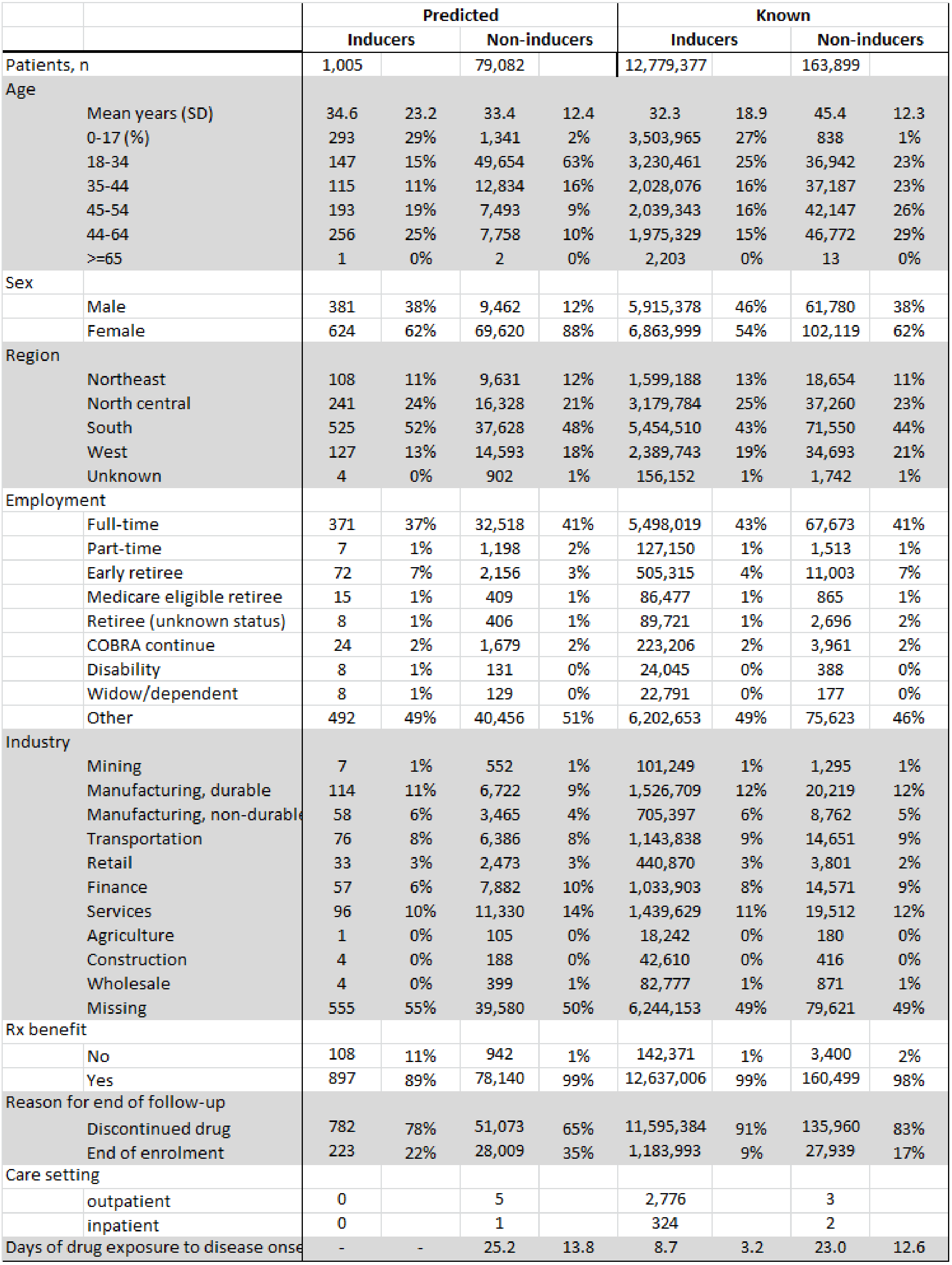
(only the first 10 rows of over 4000 rows are shown; table available upon request)



DrugBank ID	Name	predfold0	predfold1	predfold2	predfold3	predfold4	predmean	predSD
DB01581	Sulfamerazine	0.975	0.952	0.969	0.985	0.981	0.972	0.013
DB00891	Sulfapyridine	0.968	0.954	0.973	0.979	0.984	0.972	0.012
DB00576	Sulfamethizole	0.964	0.945	0.957	0.964	0.978	0.962	0.012
DB01332	Ceftizoxime	0.965	0.976	0.957	0.928	0.968	0.959	0.019
DB01325	Quinethazone	0.924	0.928	0.964	0.970	0.958	0.949	0.021
DB01298	Sulfacytine	0.918	0.909	0.935	0.964	0.952	0.936	0.023
DB01333	Cefradine	0.917	0.971	0.884	0.946	0.949	0.933	0.034
DB03294	1-Methyl-3-Oxo-1,3-Dihydro-Benzo[C]Isothiazole-5-Sulfonic Acid Amide	0.919	0.899	0.941	0.948	0.956	0.933	0.023
DB00880	Chlorothiazide	0.936	0.900	0.936	0.946	0.933	0.930	0.018
DB00689	Cephaloglycin	0.948	0.945	0.925	0.880	0.941	0.928	0.028

5 models corresponding to 5-fold CV random forest of Dragon descriptors (Drg-RF) were used for prediction.
"predfold0", "predfold1", etc. are the predictions given by models from fold0, fold1, etc.
"predmean" is the mean predicted value across the 5 models
"predSD" is the standard deviation of the predicted value across the 5 models
"NA" denotes invalid prediction (out of AD)

Table A1.4.S5. Baseline characteristics of patients employed for pharmacoepidemiology analysis



		Predicted				Known			
		Inducers		Non-inducers		Inducers		Non-inducers	
Patients, n		1,005		79,082		12,779,377		163,899	
Age									
	Mean years (SD)	34.6	23.2	33.4	12.4	32.3	18.9	45.4	12.3
	0-17 (%)	293	29%	1,341	2%	3,503,965	27%	838	1%
	18-34	147	15%	49,654	63%	3,230,461	25%	36,942	23%
	35-44	115	11%	12,834	16%	2,028,076	16%	37,187	23%
	45-54	193	19%	7,493	9%	2,039,343	16%	42,147	26%
	44-64	256	25%	7,758	10%	1,975,329	15%	46,772	29%
	>=65	1	0%	2	0%	2,203	0%	13	0%
Sex									
	Male	381	38%	9,462	12%	5,915,378	46%	61,780	38%
	Female	624	62%	69,620	88%	6,863,999	54%	102,119	62%
Region									
	Northeast	108	11%	9,631	12%	1,599,188	13%	18,654	11%
	North central	241	24%	16,328	21%	3,179,784	25%	37,260	23%
	South	525	52%	37,628	48%	5,454,510	43%	71,550	44%
	West	127	13%	14,593	18%	2,389,743	19%	34,693	21%
	Unknown	4	0%	902	1%	156,152	1%	1,742	1%
Employment									
	Full-time	371	37%	32,518	41%	5,498,019	43%	67,673	41%
	Part-time	7	1%	1,198	2%	127,150	1%	1,513	1%
	Early retiree	72	7%	2,156	3%	505,315	4%	11,003	7%
	Medicare eligible retiree	15	1%	409	1%	86,477	1%	865	1%
	Retiree (unknown status)	8	1%	406	1%	89,721	1%	2,696	2%
	COBRA continue	24	2%	1,679	2%	223,206	2%	3,961	2%
	Disability	8	1%	131	0%	24,045	0%	388	0%
	Widow/dependent	8	1%	129	0%	22,791	0%	177	0%
	Other	492	49%	40,456	51%	6,202,653	49%	75,623	46%
Industry									
	Mining	7	1%	552	1%	101,249	1%	1,295	1%
	Manufacturing, durable	114	11%	6,722	9%	1,526,709	12%	20,219	12%
	Manufacturing, non-durabl	58	6%	3,465	4%	705,397	6%	8,762	5%
	Transportation	76	8%	6,386	8%	1,143,838	9%	14,651	9%
	Retail	33	3%	2,473	3%	440,870	3%	3,801	2%
	Finance	57	6%	7,882	10%	1,033,903	8%	14,571	9%
	Services	96	10%	11,330	14%	1,439,629	11%	19,512	12%
	Agriculture	1	0%	105	0%	18,242	0%	180	0%
	Construction	4	0%	188	0%	42,610	0%	416	0%
	Wholesale	4	0%	399	1%	82,777	1%	871	1%
	Missing	555	55%	39,580	50%	6,244,153	49%	79,621	49%
Rx benefit									
	No	108	11%	942	1%	142,371	1%	3,400	2%
	Yes	897	89%	78,140	99%	12,637,006	99%	160,499	98%
Reason for end of follow-up									
	Discontinued drug	782	78%	51,073	65%	11,595,384	91%	135,960	83%
	End of enrolment	223	22%	28,009	35%	1,183,993	9%	27,939	17%
Care setting									
	outpatient	0		5		2,776		3	
	inpatient	0		1		324		2	
Days of drug exposure to disease onse		-	-	25.2	13.8	8.7	3.2	23.0	12.6

APPENDIX 2: SUPPLEMENTAL FIGURES

Supplemental figure for Chapter 2 (also available online at doi:10.1021/tx200148a)

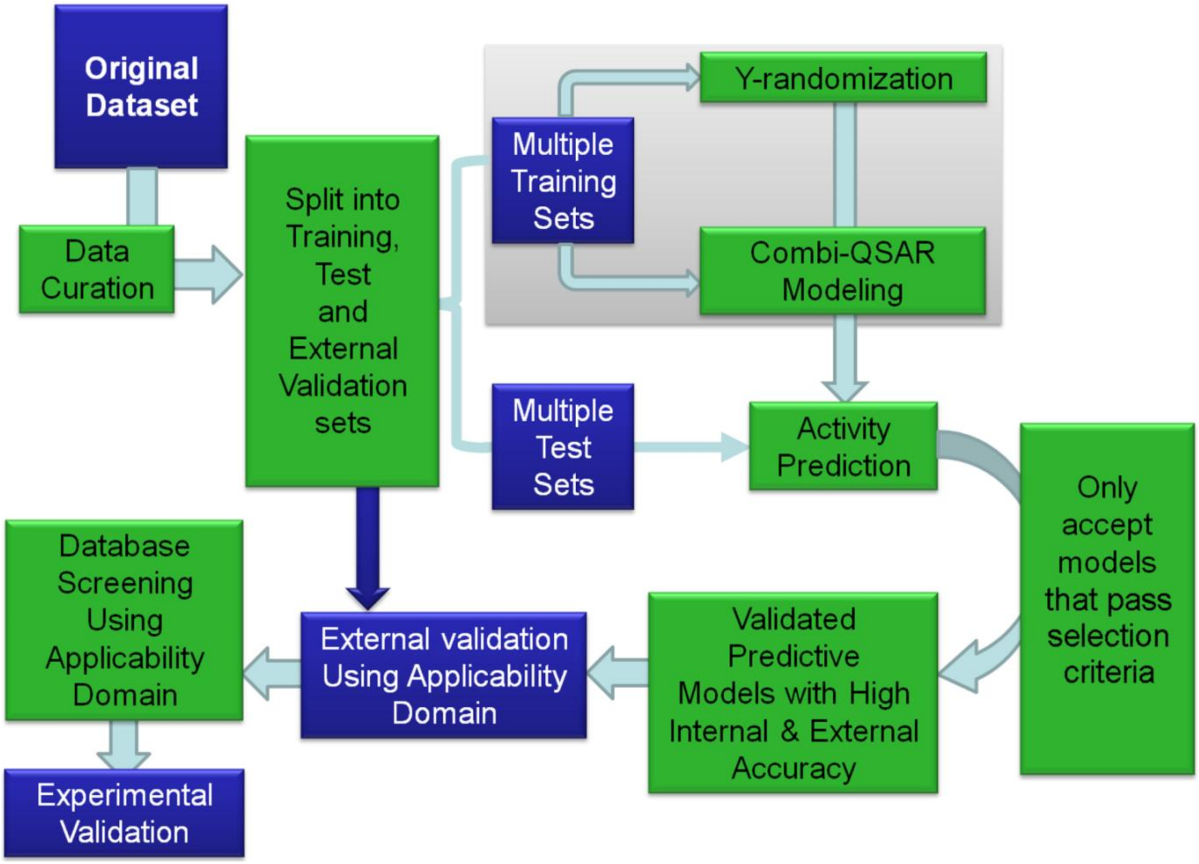

Original Dataset
Data Curation
Split into Training, Test and External Validation sets
Multiple Training Sets
Y-randomization
Combi-QSAR Modeling
Multiple Test Sets
Activity Prediction
Only accept models that pass selection criteria
Validated Predictive Models with High Internal & External Accuracy
External validation Using Applicability Domain
Database Screening Using Applicability Domain
Experimental Validation

Figure A2.2.1. Modeling workflow reproduced from Tropsha, A. (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inf. 29, 476-488

Supplemental figure for Chapter 3 (also available online at doi:10.1021/tx400110f)

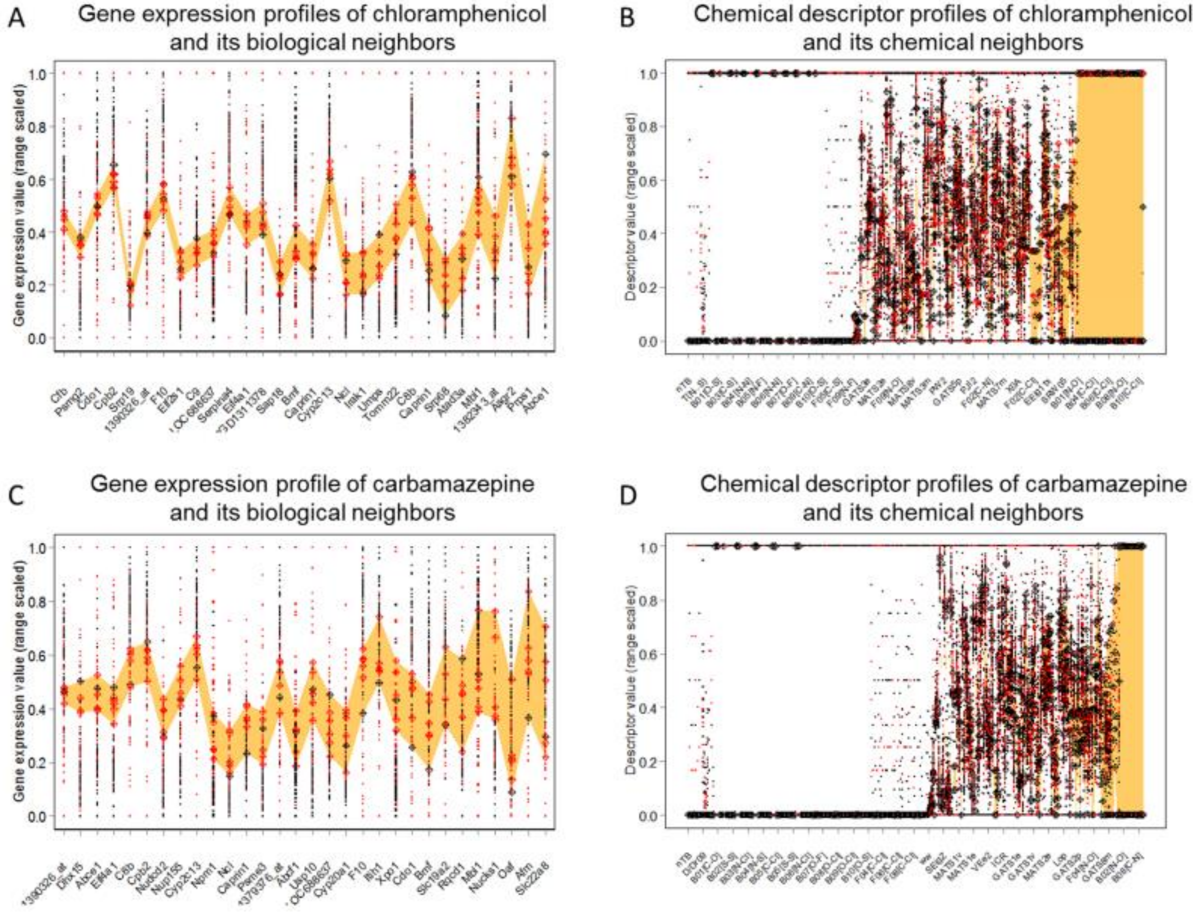


Figure A2.3.S1. Descriptor profiles of chloramphenicol and (A) its biological neighbors across 30 genes, (B), its chemical neighbors across 304 chemical descriptors, and carbamazepine and (C) its biological neighbors across 30 genes, and (D) its chemical neighbors across 304 chemical descriptors. Each diamond marks a descriptor value of the target compound or its neighbors used for RA-kNN (red for toxic, black for nontoxic). For each descriptor, these values within the neighborhood form a range (orange ribbon). Descriptors are ranked in increasing range (i.e. orange bands widen along the x-axis). Target compounds with more similar neighbors are expected to exhibit narrower orange bands (e.g. panel A shows the narrow band formed by chloramphenicol and its highly similar biological neighbors). Smaller dots mark descriptor values of all other compounds and show their distribution along each descriptor dimension. Likewise, they are colored according to toxicity (red for toxic, black for nontoxic). All 304 chemical descriptors are shown although only every other 10th descriptor is labeled on the horizontal axis.

Supplemental figures for Chapter 4

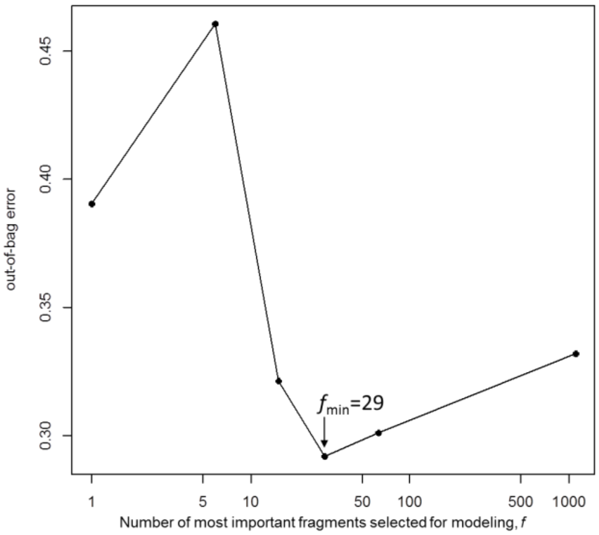


Figure A2.4.S1. Out-of-bag (OOB) error of RF model using f most important fragments. OOB error is at a minimum (0.29) when f=29 which outperformed the full model of f=1,091 fragments (OOB=0.33)

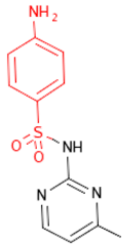
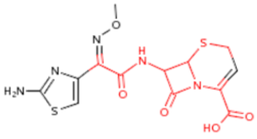
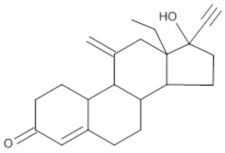
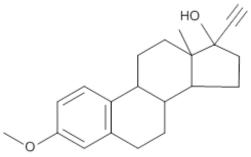
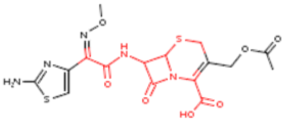
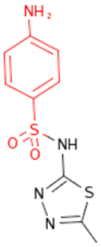
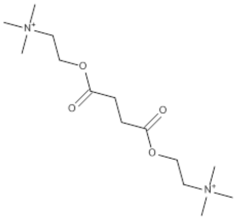
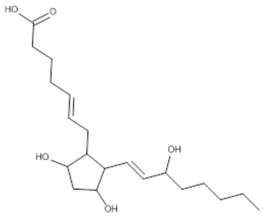
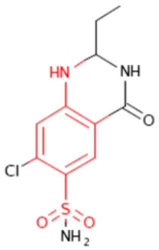
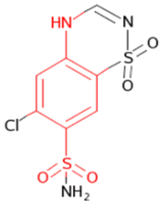
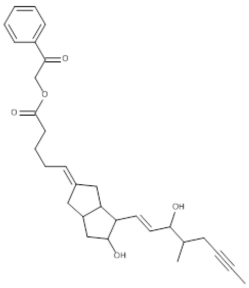
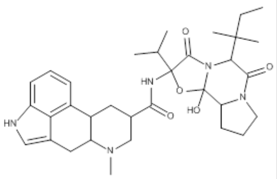
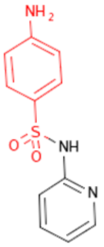
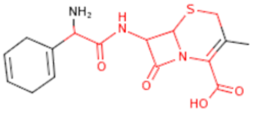
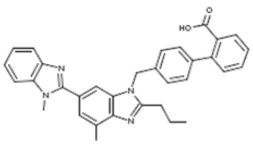
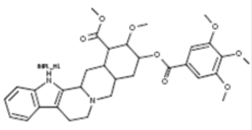
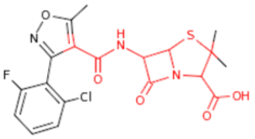
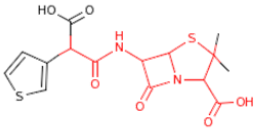
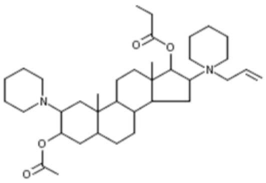
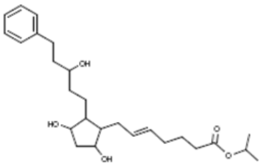
Most likely inducers		Most likely non-inducers	
 Sulfamerazine	 Ceftizoxime	 Etonogestrel	 Mestranol
 Cefotaxime	 Sulfamethizole	 Succinylcholine	 Dinoprost
 Quinethazone	 Chlorothiazide	 Iloprost	 Ergoloid mesylate
 Sulfapyridine	 Cefradine	 Telmisartan	 Deserpidine
 Flucloxacillin	 Ticarcillin	 Rapacuronium	 Latanoprost

Figure A2.4.S2. Most likely SJS inducers and non-inducers predicted by QSAR model (Dragon-RF). Structural alerts, if any, were mapped onto the predicted drugs

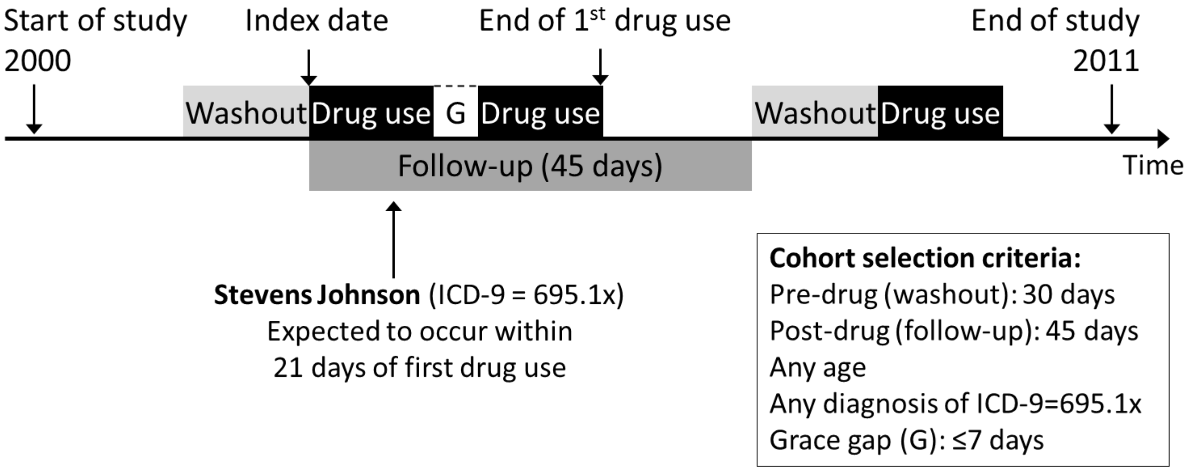


Figure A2.4.S3. Timeline and patient cohort selection criteria

APPENDIX 3: SUPPLEMENTAL METHODS

Supplemental methods for Chapter 4

Preliminary chemical exploration: SOM clustering of drugs in chemical space

To understand how the 364 drugs were distributed in the chemical space, the drugs were clustered according to their Dragon descriptor profiles using a Kohonen self-organizing map (SOM) (Kohonen 2008). Based on artificial neural networks, SOM (Kohonen package in R) projects the drugs onto a two-dimensional 6 x 6 grid of cells such that similar drugs are clustered together within a cell and similar cells are placed next to one another. Thus, the topological distance between the drugs on the SOM reflects their chemical distance from one another. The chemical distance between any two drugs is defined as the Euclidean distance of their chemical descriptor vectors.

Each cell is colored by its proportion of SJS inducers (gray if no SJS inducers, pink if all SJS inducers). Within each cell, a tag cloud of ATC codes is overlaid. The ATC letters are respectively sized and colored according to frequency and proportion of SJS inducers.

Determining p most important chemical fragments

Because the ISIDA-RF model was developed using 5-fold external CV, it involved five models which each had a slightly different RF conditional importance rank for each fragment. We used the conditional importance as… (mean diff in accuracy).

From each of the five models, the most important 10, 25, 50, 75 and 150 fragments were selected. Their intersection, made up of *p* fragments consistently among the top ranked across all five models, were selected for subsequent random forest modelling using the same

5-fold external CV scheme. The resultant reduced RF model's out-of-bag (OOB_p) error was compared to that of the full RF model, OOB_{full}. Optimal p was defined as p with the minimum OOB_p error $\leq$ OOB_{full} error. Figure S1 shows the plot of OOBp vs p where optimal p=29.

Determining significance score for co-occurring fragments.

Each co-occurring pair of fragment was tested for higher-than-expected frequency in inducers than in non-inducers by a two-tailed Fisher's exact test. As a conservative measure, p-values underwent permutation-based adjustment assuming a null distribution generated under pairwise co-occurrences with noise (randomly absent or present). Specifically for each fragment i, its pairwise co-occurrence with noise generated a null Fisher's test value, $t_{i,noise1}$. Its pairwise co-occurrences with 1000 noise fragments from 1000 permutations generated a set of test values, $t_{i,noise1}$, $t_{i,noise2}$, …, $t_{i,noise1000}$ forming the null distribution, D_i. To adjust the test value of the pairwise co-occurrence of fragments i and j, $t_{i,j}$ was compared against the relevant null distributions, D_i and D_j, such that the larger of its quantiles along the null distributions, max (q_i, q_j) was taken as the adjusted p-value. A fragment pair was said to co-occur more frequently than expected when its adjusted p-value was <0.05.

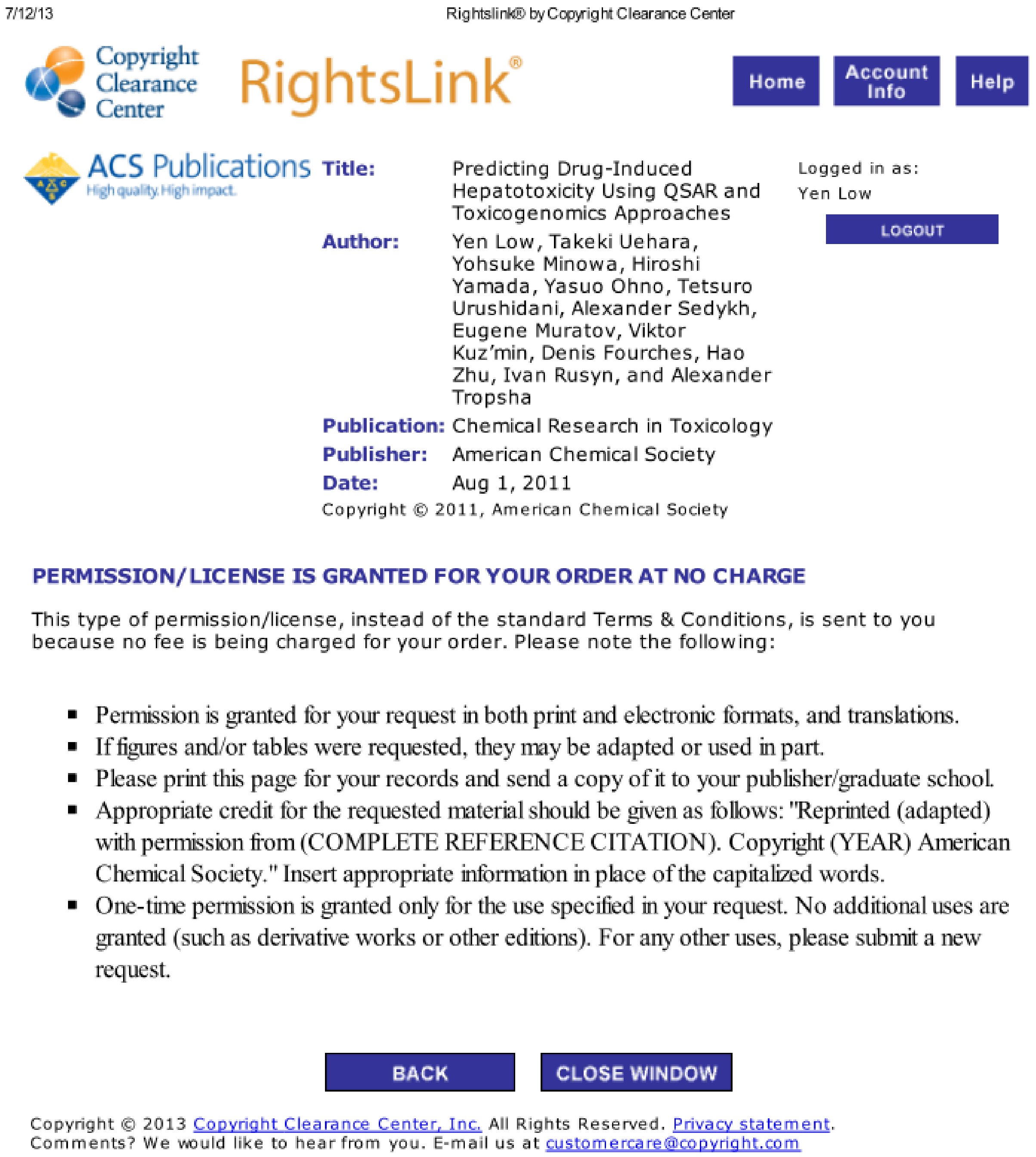
Pharmacoepidemiology

Patients using the drugs of interest were determined by outpatient prescription claims (using appropriate national drug codes, NDC) and inpatient claims (using Healthcare Common Procedure Coding System codes, HCPCS). The NDC of the drugs of interest were extracted from the FDA Redbook 2011 (according to generic name) and RxNORM (according to ingredient name). Relevant HCPCS codes were extracted if their descriptions contained the drugs of interest.

Because many of the drugs of interest (e.g. antibiotics) were used intermittently, only the first drug use period per patient fulfilling the eligibility criteria was considered. However, the same patient was included again if he/she was also eligible for the first drug use of another drug class.

APPENDIX 4: LICENSE FOR COPYRIGHTED MATERIALS

License to reproduce Chapter 2



7/12/13 Rightslink® by Copyright Clearance Center

Title:	Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches
Author:	Yen Low, Takeki Uehara, Yohsuke Minowa, Hiroshi Yamada, Yasuo Ohno, Tetsuro Urushidani, Alexander Sedykh, Eugene Muratov, Viktor Kuz'min, Denis Fourches, Hao Zhu, Ivan Rusyn, and Alexander Tropsha
Publication:	Chemical Research in Toxicology
Publisher:	American Chemical Society
Date:	Aug 1, 2011

Copyright © 2011, American Chemical Society

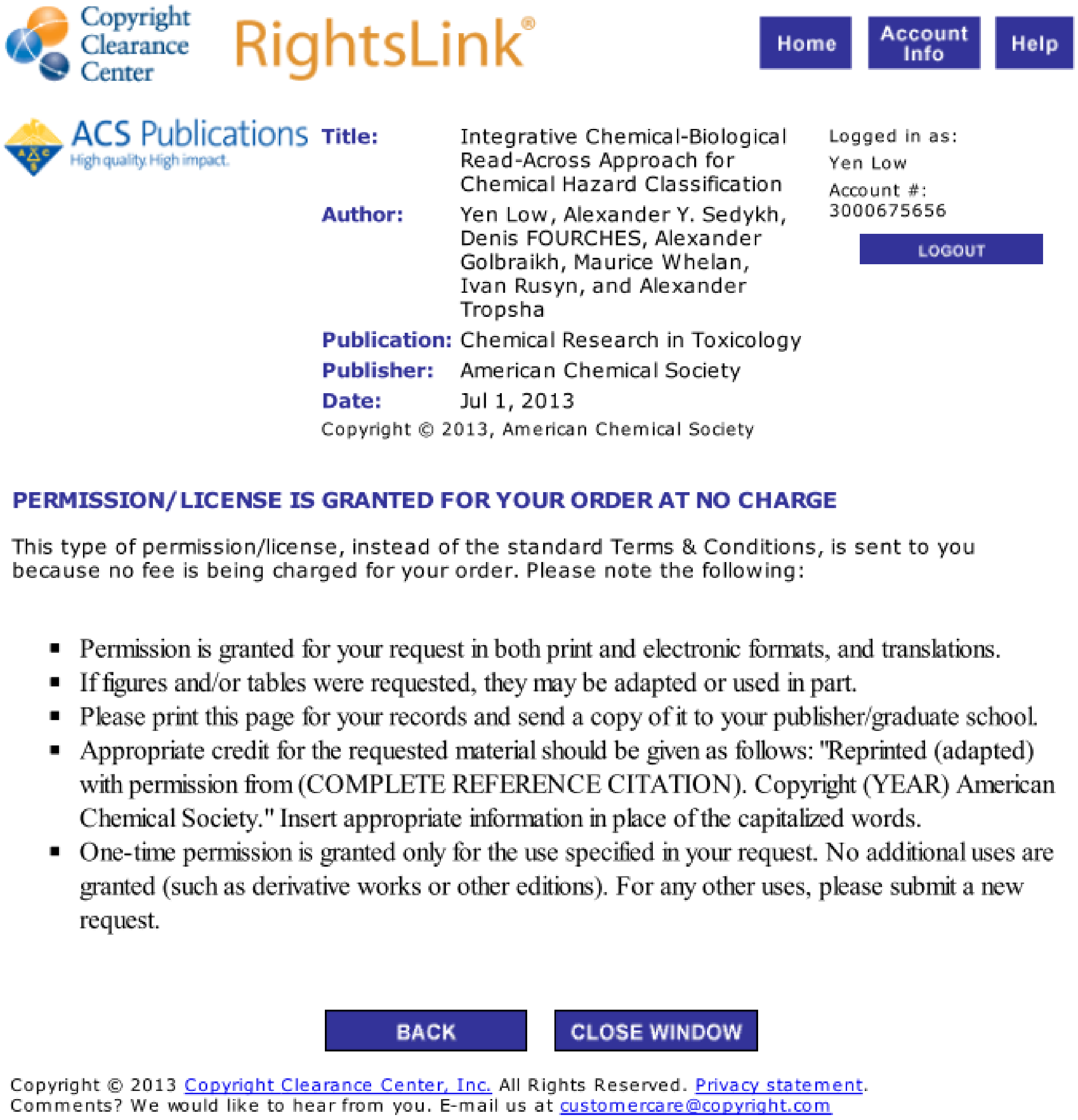
Logged in as:
Yen Low

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

License to reproduce Chapter 3



Title:	Integrative Chemical-Biological Read-Across Approach for Chemical Hazard Classification
Author:	Yen Low, Alexander Y. Sedykh, Denis FOURCHES, Alexander Golbraikh, Maurice Whelan, Ivan Rusyn, and Alexander Tropsha
Publication:	Chemical Research in Toxicology
Publisher:	American Chemical Society
Date:	Jul 1, 2013

Copyright © 2013, American Chemical Society

Logged in as:
Yen Low
Account #:
3000675656

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

REFERENCES

Aarbakke J, Bakke OM, Milde EJ, Davies DS. 1977. Disposition and oxidative metabolism of phenylbutazone in man. *Eur J Clin Pharmacol*. 11(5):359

Adam W, Ahrweiler M, Saha-Möller CR, Sauter M, Schönberger A, et al. 1993. Genotoxicity studies of benzofuran dioxetanes and epoxides with isolated DNA, bacteria and mammalian cells. *Toxicol Lett*. 67(1-3):41

Adami H-O, Berry SCL, Breckenridge CB, Smith LL, Swenberg JA, et al. 2011. Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. *Toxicol Sci*. 122(2):223

Afshari CA, Hamadeh HK, Bushel PR. 2011. The evolution of bioinformatics in toxicology: advancing toxicogenomics. *Toxicol Sci*. 120 Suppl:S225

Aha DW, Kibler D, Albert MK. 1991. Instance-based learning algorithms. *Machine Learning*. 6(1):37

Ahmad SR. 2003. Adverse drug event monitoring at the food and drug administration. *Journal of General Internal Medicine*. 18(1):57

Arellano FM. 2005. The withdrawal of rofecoxib. *Pharmacoepidemiol Drug Saf*. 14(3):213

Ashby J. 1978. Structural analysis as a means of predicting carcinogenic potential. *Br J Cancer*. 37(6):904

Austin CP, Brady LS, Insel TR, Collins FS. 2004. NIH molecular libraries initiative. *Science*. 306(5699):1138

Bai JPF, Abernethy DR. 2013. Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annu Rev Pharmacol Toxicol*. 53:451

Baker NC. 2010. *Methods in Literature-Based Drug Discovery*. University of North Carolina

Baker NC, Hemminger BM. 2010. Mining connections between chemicals, proteins, and diseases extracted from medline annotations. *J Biomed Inform*. 43(4):510

Bate A, Evans SJW. 2009. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol Drug Saf*. 18(6):427

Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, et al. 1998. A bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol*. 54(4):315

Bayada DM, Hamersma H, van Geerestein VJ. 1999. Molecular diversity and representativity in chemical databases. *J Chem Inf Comput Sci*. 39(1):1

Bender A, Scheiber J, Glick M, Davies JW, Azzaoui K, et al. 2007. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*. 2(6):861

Benigni R. 2005. Structure-activity relationship studies of chemical mutagens and carcinogens: mechanistic investigations and prediction approaches. *Chem Rev*. 105(5):1767

Berger SI, Iyengar R. 2009. Network analyses in systems pharmacology. *Bioinformatics*. 25(19):2466

Beyer RP, Fry RC, Lasarev MR, McConnachie LA, Meira LB, et al. 2007. Multicenter study of acetaminophen hepatotoxicity reveals the importance of biological endpoints in genomic analyses. *Toxicol Sci*. 99(1):326

Bisgin H, Liu Z, Fang H, Xu X, Tong W. 2011. Mining FDA drug labels using an unsupervised learning technique--topic modeling. *BMC Bioinformatics*. 12(Suppl 10):S11

Blomme EAG, Yang Y, Waring JF. 2009. Use of toxicogenomics to understand mechanisms of drug-induced hepatotoxicity during drug discovery and development. *Toxicol Lett*. 186(1):22

Brackett CC, Singh H, Block JH. 2004. Likelihood and mechanisms of cross-allergenicity between sulfonamide antibiotics and other drugs containing a sulfonamide functional group. *Pharmacotherapy*. 24(7):856

Breiman L. 2001. Random forests. *Machine Learning*. 45:5

Cami A, Arnold A, Manzi S, Reis B. 2011. Predicting adverse drug events using pharmacological network models. *Sci Transl Med*. 3(114):114ra127

Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. 2008. Drug target identification using side-effect similarity. *Science*. 321(5886):263

Caruana R. 1997. Multitask learning. *Machine Learning*. 28:41

Caster O, Norén GN, Madigan D, Bate A. 2010. Large-scale regression-based pattern discovery: the example of screening the who global drug safety database. *Statistical Analysis and Data Mining*. 3:197

Chakravarti SK, Saiakhov RD, Klopman G. 2012. Optimizing predictive performance of case ultra expert system models using the applicability domains of individual toxicity alerts. *J Chem Inf Model*. 52(10):2609

Chan HL, Stern RS, Arndt KA, Langlois J, Jick SS, et al. 1990. The incidence of erythema multiforme, Stevens-Johnson syndrome, and toxic epidermal necrolysis. a population-based study with particular reference to reactions caused by drugs among outpatients. *Arch Dermatol*. 126(1):43

Chen M, Zhang M, Borlak J, Tong W. 2012. A decade of toxicogenomic research and its contribution to toxicological science. *Toxicol Sci*. 130(2):217

Cheng F, Li W, Wang X, Zhou Y, Wu Z, et al. 2013. Adverse drug events: database construction and in silico prediction. *J Chem Inf Model*. 53(4):744

Cheng F, Liu C, Jiang J, Lu W, Li W, et al. 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*. 8(5):e1002503

Chiang AP, Butte AJ. 2009. Data-driven methods to discover molecular determinants of serious adverse drug events. *Clin Pharmacol Ther*. 85(3):259

Collander R, Lindholm M, Haug CM, Stene J, Sörensen NA. 1951. The partition of organic compounds between higher alcohols and water. *Acta Chemica Scandinavica*. 5:774

Collins FS, Gray GM, Bucher JR. 2008. Transforming environmental health protection. *Science*. 319(5865):906

Coloma PM, Trifirò G, Schuemie MJ, Gini R, Herings R, et al. 2012. Electronic healthcare databases for active drug safety surveillance: is there enough leverage? *Pharmacoepidemiol Drug Saf*. 21(6):611

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 10(6):392

Craig A, Sidaway J, Holmes E, Orton T, Jackson D, et al. 2006. Systems toxicology: integrated genomic, proteomic and metabonomic analysis of methapyrilene induced hepatotoxicity in the rat. *J Proteome Res*. 5(7):1586

Cui Y, Paules RS. 2010. Use of transcriptomics in understanding mechanisms of drug-induced toxicity. *Pharmacogenomics*. 11(4):573

Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, et al. 2010. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 26(9):1205

Dietterich TG. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15. Berlin, Heidelberg: Springer

Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci*. 95(1):5

Durant JL, Leland BA, Henry DR, Nourse JG. 2002. Reoptimization of mdl keys for use in drug discovery. *J Chem Inf Comput Sci*. 42(6):1273

Efron B, Tibshirani R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*. 1(1):54

Eisenberg DF, Daniel GW, Jones JK, Goehring EL, Wahl PM, et al. 2012. Validation of a claims-based diagnostic code for Stevens-Johnson syndrome in a commercially insured population. *Pharmacoepidemiol Drug Saf*. 21(7):760

Elder JF. 2003. The generalization paradox of ensembles. *Journal of Computational and Graphical Statistics*. 12(4):853

Enoch SJ, Cronin MTD, Schultz TW, Madden JC. 2008. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via michael addition. *Chem Res Toxicol*. 21(2):513

Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based qsars. *Environ Health Perspect*. 111(10):1361

Farombi EO, Adaramoye OA, Emerole GO. 2002. Influence of chloramphenicol on rat hepatic microsomal components and biomarkers of oxidative stress: protective role of antioxidants. *Pharmacol Toxicol*. 91(3):129

Felter SP, Vassallo JD, Carlton BD, Daston GP. 2006. A safety assessment of coumarin taking into account species-specificity of toxicokinetics. *Food Chem Toxicol*. 44(4):462

Fielden MR, Brennan R, Gollub J. 2007. A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. *Toxicol Sci*. 99(1):90

Fielden MR, Eynon BP, Natsoulis G, Jarnagin K, Banas D, Kolaja KL. 2005. A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity. *Toxicol Pathol*. 33(6):675

Fourches D, Muratov E, Tropsha A. 2010. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model*. 50(7):1189

Gatnik MF, Worth A. 2010. Review of software tools for toxicity prediction. Luxembourg

Gleeson MP, Modi S, Bender A, Robinson RLM, Kirchmair J, et al. 2012. The challenges involved in modeling toxicity data in silico: a review. *Curr Pharm Des*. 18(9):1266

Golbraikh A. 2000. Molecular dataset diversity indices and their applications to comparison of chemical databases and QSAR analysis. *J Chem Inf Comput Sci*. 40(2):414

Golbraikh A, Tropsha A. 2002a. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers*. 5(4):357

Golbraikh A, Tropsha A. 2002b. Beware of q2! *J Mol Graph Model*. 20(4):269

Guha R, Serra JR, Jurs PC. 2004. Generation of QSAR sets with a self-organizing map. *J Mol Graph Model*. 23(1):1

Hall GC, Sauer B, Bourke A, Brown JS, Reynolds MW, et al. 2012. Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf*. 21(1):1

Hällgren J, Tengvall-Linder M, Persson M, Wahlgren C. 2003. Stevens-Johnson syndrome associated with ciprofloxacin: a review of adverse cutaneous events reported in sweden as associated with this drug. *J Am Acad Dermatol*. 49(5 Suppl):S267

Hammett LP. 1937. The effect of structure upon the reactions of organic compounds. benzene derivatives. *J Am Chem Soc*. 343(1936):96

Handoko KB, van Puijenbroek EP, Bijl AH, Hermens W a JJ, Zwart-van Rijkom JEF, et al. 2008. Influence of chemical structure on hypersensitivity reactions induced by antiepileptic drugs: the role of the aromatic ring. *Drug Saf*. 31(8):695

Hansch C, Maloney P, Fujita T, Muir R. 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*. 194(4824):178

Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. 2013. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin Pharmacol Ther*. 93(6):539

Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther*. 91(6):1010

Heijne WHM, Kienhuis AS, van Ommen B, Stierum RH, Groten JP. 2005. Systems toxicology: applications of toxicogenomics, transcriptomics, proteomics and metabolomics in toxicology. *Expert Rev Proteomics*. 2(5):767

Hennessy S. 2006. Use of health care databases in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol*. 98(3):311

Hewitt M, Cronin MTD, Madden JC, Rowe PH, Johnson C, et al. 2007. Consensus QSAR models: do the benefits outweigh the complexity? *J Chem Inf Model*. 47(4):1460

Hewitt M, Ellison CM, Enoch SJ, Madden JC, Cronin MTD. 2010. Integrating (q)SAR models, expert systems and read-across approaches for the prediction of developmental toxicity. *Reprod Toxicol*. 30(1):147

Hirode M, Horinouchi A, Uehara T, Ono A, Miyagishima T, et al. 2009a. Gene expression profiling in rat liver treated with compounds inducing elevation of bilirubin. *Hum Exp Toxicol*. 28(4):231

Hirode M, Omura K, Kiyosawa N, Uehara T, Shimuzu T, et al. 2009b. Gene expression profiling in rat liver treated with various hepatotoxic-compounds inducing coagulopathy. *J Toxicol Sci*. 34(3):281

Hirode M, Ono A, Miyagishima T, Nagao T, Ohno Y, Urushidani T. 2008. Gene expression profiling in rat liver treated with compounds inducing phospholipidosis. *Toxicol Appl Pharmacol*. 229(3):290

Hou T, Wang J. 2008. Structure-ADME relationship: still a long way to go? *Expert Opin Drug Metab Toxicol*. 4(6):759

Iskar M, Zeller G, Zhao X-M, van Noort V, Bork P. 2012. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr Opin Biotechnol*. 23(4):609

Jewell NP. 1986. On the bias of commonly used measures of association for 2 x 2 tables. *Biometrics*. 42(2):351

Ji C, Kaplowitz N. 2006. ER stress: can the liver cope? *J Hepatol*. 45(2):321

Johnson MA, Maggiora GM. 1990. *Concepts and Applications of Molecular Similarity*. New York: John Wiley & Sons

Judson R, Richard A, Dix DJ, Houck K, Martin M, et al. 2009. The toxicity data landscape for environmental chemicals. *Environ Health Perspect*. 117(5):685

Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, et al. 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect*. 118(4):485

Judson RS, Kavlock RJ, Setzer RW, Cohen Hubal EA, Martin MT, et al. 2011. Estimating toxicity-related biological pathway altering doses for high-throughput chemical risk assessment. *Chem Res Toxicol*. 24(4):451

Judson RS, Martin MT, Egeghy PP, Gangwal S, Reif DM, et al. 2012. Aggregating data for computational toxicology applications: the U.S. Environmental Protection Agency (EPA) aggregated computational toxicology resource (ACToR) system. *International Journal of Molecular Sciences.* 13(2):1805

Kaufmann P, Török M, Hänni A, Roberts P, Gasser R, Krähenbühl S. 2005. Mechanisms of benzarone and benzbromarone-induced hepatic toxicity. *Hepatology*. 41(4):925

Keller DA, Juberg DR, Catlin N, Farland WH, Hess FG, et al. 2012. Identification and characterization of adverse effects in 21st century toxicology. *Toxicol Sci*. 126(2):291

Kienhuis AS, van de Poll MCG, Wortelboer H, van Herwijnen M, Gottschalk R, et al. 2009. Parallelogram approach using rat-human in vitro and rat in vivo toxicogenomics predicts acetaminophen-induced hepatotoxicity in humans. *Toxicol Sci*. 107(2):544

King DE, Malone R, Lilley SH. 2000. New classification and update on the quinolone antibiotics. *Am Fam Physician*. 61(9):2741

Kiyosawa N, Uehara T, Gao W, Omura K, Hirode M, et al. 2007. Identification of glutathione depletion-responsive genes using phorone-treated rat liver. *The Journal of Toxicological Sciences*. 32(5):469

Kohonen T. 2008. *Self Organizing Maps*. Heidelberg: Springer. 3rd ed.

Komar CM. 2005. Peroxisome proliferator-activated receptors (ppars) and ovarian function--implications for regulating steroidogenesis, differentiation, and tissue remodeling. *Reprod Biol Endocrinol*. 3:41

Kovatcheva A, Golbraikh A, Oloff S, Xiao Y-D, Zheng W, et al. 2004. Combinatorial QSAR of ambergris fragrance compounds. *J Chem Inf Comput Sci*. 44(2):582

Kruhlak NL, Benz RD, Zhou H, Colatsky TJ. 2012. (q)SAR modeling and safety assessment in regulatory review. *Clin Pharmacol Ther*. 91(3):529

Kunishima C, Inoue I, Oikawa T, Nakajima H, Komoda T, Katayama S. 2007. Activating effect of benzbromarone, a uricosuric drug, on peroxisome proliferator-activated receptors. *PPAR Res*. 2007:36092

Kuz'min VE, Artemenko AG, Muratov EN. 2008. Hierarchical QSAR technology based on the simplex representation of molecular structure. *J Comput Aided Mol Des*. 22(6-7):403

Kuz'min VE, Polishchuk PG, Artemenko AG, Andronati S a. 2011. Interpretation of QSAR models based on random forest methods. *Molecular Informatics*. 30(6-7):593

Lake BG, Gray TJB, Evans JG, Lewis DFV, Beamand JA, Hue KL. 1989. Studies on the mechanism of coumarin-induced toxicity in rat hepatocytes: comparison with dihydrocoumarin and other coumarin metabolites. *Toxicology and Applied Pharmacology*. 97(2):311

Lee M-HH, Graham GG, Williams KM, Day RO. 2008. A benefit-risk assessment of benzbromarone in the treatment of gout. was its withdrawal from the market in the best interest of patients? *Drug Saf*. 31(8):643

Lee PW, Neal RA. 1978. Metabolism of methimazole by rat liver cytochrome P-450-containing monoxygenases. *Drug Metab Dispos*. 6(5):591

LePendu P, Iyer S V, Bauer-Mehren A, Harpaz R, Mortensen JM, et al. 2013. Pharmacovigilance using clinical notes. *Clin Pharmacol Ther*. 93(6):547

Liang L, Cai F, Cherkassky V. 2009. Predictive learning with structured (grouped) data. *Neural Netw*. 22(5-6):766

Lin C-J, Malina A, Pelletier J. 2009. c-myc and eif4f constitute a feedforward loop that regulates cell growth: implications for anticancer therapy. *Cancer Res*. 69(19):7491

Lindquist M. 2008. VigiBase, the WHO global ICSR database system: basic facts. *Drug Information Journal*. 42:409

Lipkind GM, Fozzard HA. 2010. Molecular model of anticonvulsant drug binding to the voltage-gated sodium channel inner pore. *Mol Pharmacol*. 78(4):631

Liu M, Wu Y, Chen Y, Sun J, Zhao Z, et al. 2012. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc*. 19(e1):e28

Lock EF, Abdo N, Huang R, Xia M, Kosyk O, et al. 2012. Quantitative high-throughput screening for chemical toxicity in a population-based in vitro model. *Toxicol Sci*. 126(2):578

Loew GH, Goldblum A. 1985. Metabolic activation and toxicity of acetaminophen and related analogs. a theoretical study. *Mol Pharmacol*. 27(3):375

Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, et al. 2012. Large-scale prediction and testing of drug activity on side-effect targets. *Nature*. 486(7403):361

Lounkine E, Nigsch F, Jenkins JL, Glick M. 2011. Activity-aware clustering of high throughput screening data and elucidation of orthogonal structure-activity relationships. *J Chem Inf Model*. 51(12):3158

Low Y, Sedykh AY, Fourches D, Golbraikh A, Whelan M, et al. 2013a. Integrative chemical-biological read-across approach for chemical hazard classification. *Chem Res Toxicol*. Epup July

Low Y, Sedykh A, Golbraikh A, Tropsha A, Fourches D. 2013b. Chemistry-wide association studies (CWAS): determining joint mutagenic effects of co-occurring chemical features

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(8):1251

Luebke-Wheeler J, Zhang K, Battle M, Si-Tayeb K, Garrison W, et al. 2008. Hepatocyte nuclear factor 4alpha is implicated in endoplasmic reticulum stress-induced acute phase response by regulating expression of cyclic adenosine monophosphate responsive element binding protein h. *Hepatology*. 48(4):1242

Maggiora GM. 2006. On outliers and activity cliffs--why QSAR often disappoints. *J Chem Inf Model*. 46(4):1535

Marron JS, Todd MJ, Ahn J. 2007. Distance-weighted discrimination. *J Am Stat Assoc*. 102(480):1267

Matthews EJ, Kruhlak NL, Daniel Benz R, Sabaté DA, Marchant CA, Contrera JF. 2009a. Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans: part C: use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities. *Regul Toxicol Pharmacol*. 54(1):43

Matthews EJ, Ursem CJ, Kruhlak NL, Benz RD, Sabaté DA, et al. 2009b. Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: part B. use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Regul Toxicol Pharmacol*. 54(1):23

McClure DL, Raebel MA, Yih WK, Mullersman J, Anderson-Smits C, et al. 2012. Mini-Sentinel methods: framework for assessment of signal refinement positive results

Merlot C. 2008. In silico methods for early toxicity assessment. *Curr Opin Drug Discov Devel*. 11(1):80

Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*. 26(4):445

Mortelmans K, Zeiger E. 2000. The Ames salmonella/microsome mutagenicity assay. *Mutat Res*. 455(1-2):29

Muratov EN, Artemenko AG, Varlamova E V, Polischuk PG, Lozitsky VP, et al. 2010. Per aspera ad astra: application of simplex QSAR approach in antiviral research. *Future Medicinal Chemistry*. 2(7):1205

Naisbitt DJ, Hough SJ, Gill HJ, Pirmohamed M, Kitteringham NR, Park BK. 1999. Cellular disposition of sulphamethoxazole and its metabolites: implications for hypersensitivity. *Br J Pharmacol*. 126(6):1393

National Academy of Sciences, National Research Council. 2007. *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington DC: National Academies Press

National Research Council. 2007. *Preventing medication errors: quality chasm series*. Washington DC: National Academies Press

Natsoulis G, Pearson CI, Gollub J, P Eynon B, Ferng J, et al. 2008. The liver pharmacological and xenobiotic gene response repertoire. *Mol Syst Biol*. 4(175):175

Nikolova N, Jaworska J. 2003. Approaches to measure chemical similarity– a review. *QSAR & Combinatorial Science*. 22(910):1006

Norén GN, Hopstadius J, Bate A. 2011. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. *Stat Methods Med Res*

OECD. 2007. Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. *69*, Paris

Oliveira JL, Lopes P, Nunes T, Campos D, Boyer S, et al. 2012. The EU-ADR web platform: delivering advanced pharmacovigilance tools. *Pharmacoepidemiol Drug Saf*

Oprea TI, Nielsen SK, Ursu O, Yang JJ, Taboureau O, et al. 2011. Associating drugs, targets and clinical outcomes into an integrated network affords a new platform for computer-aided drug repurposing. *Molecular Informatics*. 30(2-3):100

Oprea TI, Tropsha A, Faulon J-L, Rintoul MD. 2007. Systems chemical biology. *Nat Chem Biol*. 3(8):447

Parviz F, Matullo C, Garrison WD, Savatski L, Adamson JW, et al. 2003. Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nat Genet*. 34(3):292

Patel CJ, Bhattacharya J, Butte AJ. 2010. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*. 5(5):e10746

Pauwels E, Stoven V, Yamanishi Y. 2011. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*. 12(1):169

Pearlman R, Smith K. 1998. Novel software tools for chemical diversity. In *Perspectives in Drug Discovery and Design*. 2:339. Kluwer Academic Publishers

Peters JM, Morishima H, Ward JM, Coakley CJ, Kimura S, Gonzalez FJ. 1999. Role of CYP1A2 in the toxicity of long-term phenacetin feeding in mice. *Toxicol Sci*. 50(1):82

Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, et al. 2012. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf*. 21 Suppl 1:1

Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, et al. 2009. Application of random forest approach to QSAR prediction of aquatic toxicity. *J Chem Inf Model*. 49(11):2481

Pons P, Latapy M. 2005. Computing communities in large networks using random walks. , pp. 1–20

Porter WR, Neal RA. 1978. Metabolism of thioacetamide and thioacetamide s-oxide by rat liver microsomes. *Drug Metab Dispos*. 6(4):379

Pouliot Y, Chiang AP, Butte AJ. 2011. Predicting adverse drug reactions using publicly available PubChem bioassay data. *Clin Pharmacol Ther*. 90(1):90

Ray W a. 2003. Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology*. 158(9):915

Rechnagel RO, Glende EA. 1973. Carbon tetrachloride hepatotoxicity: an example of lethal cleavage. *CRC Crit Rev Toxicol*. 2(3):263

Reif DM, Sypa M, Lock EF, Wright F a, Wilson A, et al. 2013. ToxPi GUI: an interactive visualization tool for transparent integration of data from diverse sources of evidence. *Bioinformatics*. 29(3):402

Reilly TP, Ju C. 2002. Mechanistic perspectives on sulfonamide-induced cutaneous drug reactions. *Curr Opin Allergy Clin Immunol*. 2(4):307

Rodgers AD, Zhu H, Fourches D, Rusyn I, Tropsha A. 2010. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem Res Toxicol*. 23(4):724

Rosenkranz H, Klopman G. 1988. CASE, the computer-automated structure evaluation system, as an alternative to extensive animal testing. *Toxicol Ind Health*. 4(4):533

Rotroff DM, Wetmore B a, Dix DJ, Ferguson SS, Clewell HJ, et al. 2010. Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. *Toxicol Sci*. 117(2):348

Roujeau JC, Kelly JP, Naldi L, Rzany B, Stern RS, et al. 1995. Medication use and the risk of Stevens-Johnson syndrome or toxic epidermal necrolysis. *N Engl J Med*. 333(24):1600

Roujeau J-C, Bricard G, Nicolas J-F. 2011. Drug-induced epidermal necrolysis: important new piece to end the puzzle. *J Allergy Clin Immunol*. 128(6):1277

Rusyn I, Sedykh A, Low Y, Guyton KZ, Tropsha A. 2012. Predictive modeling of chemical hazard by integrating numerical descriptors of chemical structures and short-term toxicity assay data. *Toxicol Sci*. 127(1):1

Scheiber J, Jenkins JL, Sukuru SCK, Bender A, Mikhailov D, et al. 2009. Mapping adverse drug reactions in chemical space. *J Med Chem*. 52(9):3103

Schneider G, Kachroo S, Jones N, Crean S, Rotella P, et al. 2012. A systematic review of validated methods for identifying erythema multiforme major/minor/not otherwise specified, Stevens-Johnson syndrome, or toxic epidermal necrolysis using administrative and claims data. *Pharmacoepidemiol Drug Saf*. 21 Suppl 1:236

Schuster D, Laggner C, Langer T. 2005. Why drugs fail--a study on side effects in new chemical entities. *Curr Pharm Des*. 11(27):3545

Schüürmann G, Ebert R-U, Kühne R. 2011. Quantitative read-across for predicting the acute fish toxicity of organic compounds. *Environ Sci Technol*. 45(10):4616

Scior T, Medina-Franco JL, Do Q-T, Martínez-Mayorga K, Yunes Rojas J a, Bernard P. 2009. How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr Med Chem*. 16(32):4297

Sedykh A, Zhu H, Tang H, Zhang L, Richard A, et al. 2011. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ Health Perspect*. 119(3):364

Sedykh AY, Klopman G. 2006. A structural analogue approach to the prediction of the octanol-water partition coefficient. *J Chem Inf Model*. 46(4):1598

Selassie C, Verma RP. 2010. History of quantitative structure–activity relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, ed. DJ Abraham. 1:1. Hoboken, NJ, USA: John Wiley & Sons, Inc. 7th ed.

Shen ML, Johnson KL, Mays DC, Lipsky JJ, Naylor S. 2001. Determination of in vivo adducts of disulfiram with mitochondrial aldehyde dehydrogenase. *Biochem Pharmacol*. 61(5):537

Shetty KD, Dalal SR. 2011. Using information mining of the medical literature to improve drug safety. *J Am Med Inform Assoc*. 18(5):668

Shirakuni Y, Okamoto K, Uejima E, Inui S, Takahara J -i., et al. 2012. A practical estimation method for analyzing adverse drug reactions using data mining. *Drug Information Journal*. 47(2):235

Shmueli G. 2010. To explain or to predict? *Statistical Science*. 25(3):289

Sridhar SK, Pandeya SN, Stables JP, Ramesh A. 2002. Anticonvulsant activity of hydrazones, schiff and mannich bases of isatin derivatives. *Eur J Pharm Sci*. 16(3):129

Stephenson J. 2009. Chemical regulation: observations on improving the Toxic Substances Control Act. Washington DC

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*. 9:307

Strom BL, Carson JL, Halpern AC, Schinnar R, Snyder ES, et al. 1991a. A population-based study of Stevens-Johnson syndrome. incidence and antecedent drug exposures. *Arch Dermatol*. 127(6):831

Strom BL, Carson JL, Halpern AC, Schinnar R, Snyder ES, et al. 1991b. Using a claims database to investigate drug-induced Stevens-Johnson syndrome. *Statistics in Medicine*. 10(4):565

Strom BL, Kimmel SE, Hennessy S. 2012. *Pharmacoepidemiology*. Hongkong: Wiley-Blackwell. 5th ed.

Swanson DR. 1986. Fish oil, Raynaud’s syndrome, and undiscovered public knowledge. *Perspect Biol Med*. 30(1):7

Taft RW. 1952. Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters. *J Am Chem Soc*. 74(11):2729

Tamura K, Ono A, Miyagishima T, Nagao T, Urushidani T. 2006. Profiling of gene expression in rat liver and rat primary cultured hepatocytes treated with peroxisome proliferators. *J Toxicol Sci*. 31(5):471

Tatonetti NP, Liu T, Altman RB. 2009. Predicting drug side-effects by chemical systems biology. *Genome Biol*. 10(9):238

Tatonetti NP, Ye PP, Daneshjou R, Altman RB. 2012. Data-driven prediction of drug effects and interactions. *Sci Transl Med*. 4(125):125ra31

Thomas RS, Black MB, Li L, Healy E, Chu T-M, et al. 2012. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicol Sci*. 128(2):398

Todeschini R, Consonni V. 2000. *Handbook of Molecular Descriptors*. Weinheim, Germany: Wiley-VCH Verlag GmbH

Toler SM, Rodriguez I. 2004. Not all sulfa drugs are created equal. *Ann Pharmacother*. 38(12):2166

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, et al. 2013. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Brief Bioinformatics*. 14(3):315

Tropsha A. 2010. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*. 29:476

Tropsha A, Golbraikh A. 2007. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des*. 13(34):3494

Tropsha A, Gramatica P, Gombar VK. 2003. The importance of being earnest: validation is the absolute essential for successful application and interpretation of qspr models. *QSAR & Combinatorial Science*. 22(1):69

Truven Health Analytics. *MarketScan*. http://www.truvenhealth.com/your_healthcare_focus/pharmaceutical_and_medical_device/data_databases_and_online_tools.aspx

Tusher VG, Tibshirani R, Chu G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*. 98(9):5116

Uehara T, Hirode M, Ono A, Kiyosawa N, Omura K, et al. 2008. A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats. *Toxicology*. 250(1):15

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(2):218

Uetrecht J. 2002. N-oxidation of drugs associated with idiosyncratic drug reactions. *Drug Metab Rev*. 34(3):651

UNECE. 2009. *Health hazards, In Globally Harmonized System of Classification and Labelling of Chemicals (GHS)*

Valerio LG, Choudhuri S. 2012. Chemoinformatics and chemical genomics: potential utility of in silico methods. *J Appl Toxicol*. 32(11):880

Vapnik VN. 2000. *The Nature of Statistical Learning Theory*. New York: Springer

Varnek A, Fourches D, Hoonakker F, Solov'ev VP. 2005. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J Comput Aided Mol Des*. 19(9-10):693

Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, et al. 2008. ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Current Computer - Aided Drug Design*. 4(3):191

Vassallo JD, Hicks SM, Daston GP, Lehman-McKeeman LD. 2004. Metabolic detoxification determines species differences in coumarin-induced hepatotoxicity. *Toxicol Sci*. 80(2):249

Vilar S, Harpaz R, Chase HS, Costanzi S, Rabadan R, Friedman C. 2011. Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis. *J Am Med Inform Assoc*. 18 Suppl 1:i73

Vilar S, Harpaz R, Santana L, Uriarte E, Friedman C. 2012. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis. *PLoS ONE*. 7(7):e41471

Voutchkova AM, Osimitz TG, Anastas PT. 2010. Toward a comprehensive molecular design framework for reduced hazard. *Chem Rev*. 110(10):5845

Wang NCY, Jay Zhao Q, Wesselkamper SC, Lambert JC, Petersen D, Hess-Wilson JK. 2012. Application of computational toxicological approaches in human health risk assessment. i. a tiered surrogate approach. *Regul Toxicol Pharmacol*. 63(1):10

Wetmore B a, Wambaugh JF, Ferguson SS, Sochaski MA, Rotroff DM, et al. 2011. Integration of dosimetry, exposure and high-throughput screening data in chemical toxicity assessment. *Toxicol Sci*. 125(1):157

White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. 2013. Web-scale pharmacovigilance: listening to signals from the crowd. *J Am Med Inform Assoc*. 20(3):404

Willett P. 2000. Chemoinformatics - similarity and diversity in chemical libraries. *Curr Opin Biotechnol*. 11(1):85

Willett P, Barnard JM, Downs GM. 1998. Chemical similarity searching. *J Chem Inf Comput Sci*. 38(6):983

Wilson AM, Thabane L, Holbrook A. 2003. Application of data mining techniques in pharmacovigilance. *Br J Clin Pharmacol*. 57(2):127

Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, et al. 2012. SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics*. 13(1):164

Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. 2008. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*. 36(Database issue):D901

Wold S, Eriksson L. 1995. Statistical validation of QSAR results. In *Chemometrics Methods in Molecular Design*, ed. H van de Waterbeemd, pp. 309–18. Weinheim (Germany): VCH

Wu S, Blackburn K, Amburgey J, Jaworska J, Federle T. 2010. A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments. *Regul Toxicol Pharmacol*. 56(1):67

Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. 2008. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*. 24(13):i232

Yamanishi Y, Pauwels E, Kotera M. 2012. Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Inf Model*. 52(12):3284

Yang L, Wang K, Chen J, Jegga AG, Luo H, et al. 2011. Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome--clozapine-induced agranulocytosis as a case study. *PLoS Comput Biol*. 7(3):e1002016

Zhang L. 2011. *Development and application of cheminformatics approaches to facilitate drug discovery and environmental toxicity assessment*. University of North Carolina at Chapel Hill

Zhang L, Sedykh A, Tripathi A, Zhu H, Afantitis A, et al. 2013. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol Appl Pharmacol*. Epub May

Zhang M, Chen M, Tong W. 2012. Is toxicogenomics a more reliable and sensitive biomarker than conventional indicators from rats to predict drug-induced liver injury in humans? *Chem Res Toxicol*. 25(1):122

Zheng W, Tropsha A. 2000. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci*. 40(1):185

Zhu H, Rusyn I, Richard A, Tropsha A. 2008. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ Health Perspect*. 116(4):506

Zhu H, Ye L, Richard A, Golbraikh A, Wright F a, et al. 2009. A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environ Health Perspect*. 117(8):1257

Zhu X-W, Sedykh A, Liu S-S. 2013. Hybrid in silico models for drug-induced liver injury using chemical descriptors and in vitro cell-imaging information. *J Appl Toxicol*. Epub 22 February

Zidek N, Hellmann J, Kramer P-JJ, Hewitt PG. 2007. Acute hepatotoxicity: a predictive model based on focused Illumina microarrays. *Toxicol Sci*. 99(1):289

Zvinavashe E, Murk AJ, Rietjens IMCM. 2008. Promises and pitfalls of quantitative structure-activity relationship approaches for predicting metabolism and toxicity. *Chem Res Toxicol*. 21(12):2229