

KNOWLEDGE-GUIDED DEEP LEARNING MODELS OF DRUG TOXICITY IMPROVE

INTERPRETATION

Yun Hao

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Li Shen

Professor of Informatics in Biostatistics

and Epidemiology

Co-Supervisor of Dissertation

Jason H. Moore

Adjunct Professor of Informatics in

Biostatistics and Epidemiology

Graduate Group Chairperson

Benjamin F. Voight

Associate Professor of Genetics and Systems Pharmacology and Translational Medicine

Dissertation Committee

Chair: John H. Holmes, Professor of Medical Informatics in Epidemiology

Trevor M. Penning, Thelma Brown and Henry Charles Molinoff Professor of Pharmacology

Nicholas P. Tatonetti, Associate Professor of Biomedical Informatics

Ryan J. Urbanowicz, Assistant Professor of Informatics

KNOWLEDGE-GUIDED DEEP LEARNING MODELS OF DRUG TOXICITY IMPROVE
INTERPRETATION

COPYRIGHT

2022

Yun Hao

This work is licensed under the
Creative Commons Attribution-
NonCommercial-ShareAlike 4.0
License

To view a copy of this license, visit

<https://creativecommons.org/licenses/by-nc-sa/4.0/us/>

To my parents, Zilong Hao and Wei Dai, who always love and support me unconditionally

ACKNOWLEDGMENT

First and foremost, I would like to express my sincere gratitude to Dr. Jason Moore. In all senses, my achievement at Penn is owing to Jason being my thesis advisor. He is a world-class researcher in the field of bioinformatics and computational biology. His vision and insight for deep learning inspired my thesis research. He is also an excellent mentor. His wisdom and expertise guided me through the ups and downs of my PhD journey. It was a great experience working with him and his research team. I am grateful for all his advises and support.

I would also like to extend the gratitude to my wonderful thesis committee members, Dr. John Holmes, Dr. Li Shen, Dr. Trevor Penning, Dr. Ryan Urbanowicz, and Dr. Nicholas Tatonetti. Specifically, I would like to thank John for taking on the role of committee chair and convening all the meetings, Li for welcoming me into his lab after Jason's departure from Penn, Trevor for accommodating me to his Molecular Toxicology course, Ryan for helping me using his expertise in feature selection, and Nick for guiding me into the field of computational toxicology and supporting me throughout this amazing journey. I am grateful for all my committee members taking their time, helping me put together my thesis research, and providing invaluable feedback from diverse perspectives: biomedical informatics, machine learning, system pharmacology, and molecular toxicology. Their dedication to mentoring always inspires me to become a better scientist.

I am extremely grateful for being part of the MooreLab—my academic home for the past four years. The MooreLab provided me with a remarkable environment to conduct my thesis research. Specifically, I would like to thank Dr. Joseph Romano. Joe and I have been working alongside each other for the past eight years, from Columbia to Penn. His collaboration makes an indispensable part of my thesis research. I would also like to thank my fellow PhD students in the lab: Alexa Woodward, John Gregg, and Van Truong, for supporting each other along our journeys.

I am also grateful for being part of the GCB community. The GCB program provided me with an excellent training environment and a comprehensive view to learn and grow as a scientist.

Specifically, I would like to thank Dr. Li-San Wang and Dr. Benjamin Voight for recruiting me and welcoming me into the program, Maureen Kirsch and Anne-Cara Apple for helping me with administrative paperwork and keeping GCB running smoothly. I would also like to give a shout out to my amazing PhD cohort: Jake Crawford, David Wang, Ariel Hippen, Ben Heil, Parisa Samareh, Matthew Gazzara, Kaylyn Clark, and John Gregg, for their firm support and friendship.

Last but not least, I am forever grateful for the unconditional love and support from my parents, who inspire and motivate me every single day. And I would not be standing here without my friends and family - past, present, and future, in Shenyang, Shanghai, New Haven, New York, Philadelphia, and all around the world. I cannot express my gratitude enough for the friendship and companion, for the adventure and fun, for the laughter and tears, for all the amazing time we have spent and will spend together: Jonah Keough, Steve MacCroy, Damien and Rafi Mittlefehldt, Jacob Garcia, Bob Putney, Jim Moran, Michelle Chau, Silis Jiang and Fernanda Polubriaginof, and so many others. This is a dedication to all of you.

ABSTRACT

KNOWLEDGE-GUIDED DEEP LEARNING MODELS OF DRUG TOXICITY IMPROVE INTERPRETATION

Yun Hao

Li Shen

Jason H. Moore

In drug development, a major reason for attrition is the lack of understanding of cellular mechanisms governing drug toxicity. Conventional models are limited by low accuracy and lack of interpretability. Further, they often fail to explain cellular mechanisms underlying structure-toxicity associations. To address these limitations, we developed a series of interpretable *in silico* models that connect drugs to their toxicity targets and pathways. In Chapter 2, we incorporated target profile as an intermediate connecting structure to toxicity. To accommodate for high-dimensional feature space, we developed a pipeline named TargetTox that can identify subset of predictive features. The features identified by TargetTox accurately predicted binding outcomes for 377 targets and toxicity outcomes for 36 adverse events. We demonstrated that predictive targets tend to be differentially expressed in the tissue of toxicity. We also discovered that predictive targets are enriched for key cellular functions associated with toxicity. Furthermore, we found evidence supporting diagnostic/therapeutic applications of some predictive targets. Our findings highlighted the critical role of predictive targets in cellular mechanisms leading to toxicity. In Chapter 3, we developed DTox, an interpretation framework for knowledge-guided neural networks, which can predict compound response to toxicity assays and infer toxicity pathways of individual compounds. We demonstrate that DTox can achieve the same level of predictive performance as conventional models with a significant improvement in interpretability. Using DTox, we were able to rediscover mechanisms of transcription activation by nuclear receptors, recapitulate cellular activities induced by aromatase inhibitors and PXR agonists, and differentiate distinctive mechanisms leading to HepG2 cytotoxicity. Virtual screening by DTox revealed that

compounds with predicted cytotoxicity are at higher risk for clinical hepatic phenotypes. In Chapter 4, we introduce AIDTox, an interpretable deep learning model which incorporates curated knowledge of chemical-gene connections, gene-pathway annotations, and pathway hierarchy. AIDTox can accurately predict cytotoxicity outcomes. It also provides comprehensive explanations of cytotoxicity covering multiple aspects of drug activity including target interaction, metabolism, and elimination. In summary, our work provides a framework for deciphering cellular mechanisms of toxicity *in silico*.

Table of Contents

ACKNOWLEDGMENT	iv
ABSTRACT	vi
LIST OF TABLES	xii
LIST OF ILLUSTRATIONS	xiii
CHAPTER 1: INTRODUCTION	1
1.1 Toxicity is a major cause of attrition in drug development	1
1.2 Toxicity can be caused by distinctive underlying cellular mechanisms	1
1.3 Large-scale toxicological data opens the door for computational research	2
1.4 Various in silico models have been developed for toxicity assessment	3
1.5 Conventional QSAR models are limited by low accuracy and lack of interpretability	5
CHAPTER 2: FEATURE SELECTION PIPELINE IDENTIFIES PREDICTIVE TARGETS ASSOCIATED WITH DRUG TOXICITY	7
2.1 Introduction	7
2.2 Materials and Methods	8
2.2.1 Building compound-target interaction datasets	8
2.2.2 Building drug-adverse event datasets	8
2.2.3 Incorporating ReBATE methods to build a feature selection pipeline	9
2.2.4 Identifying optimal setting of hyperparameters by grid search	9
2.2.5 Implementing TargetTox to identify targets predictive for adverse events	10
2.2.6 Comparing similarity of predictive descriptors between targets	10
2.2.7 Comparing predictive targets identified by TargetTox to DisGeNET	10
2.2.8 Clustering adverse events by predictive targets	11
2.2.9 Analyzing differential expression of predictive targets in tissue of toxicity	11
2.2.10 Identifying enriched GO terms for predictive targets	12
2.2.11 Identifying disease genes from predictive targets	12
2.3 Results	13
2.3.1 TargetTox can accurately predict binding outcomes for targets	13
2.3.2 TargetTox can identify similar structure properties for target proteins of similar function	15

2.3.3	<i>TargetTox can achieve same level of performance as QSAR in toxicity outcome prediction</i>	16
2.3.4	<i>Similar adverse events can be clustered together by predictive targets</i>	17
2.3.5	<i>Predictive targets are differentially expressed in the tissue of toxicity</i>	18
2.3.6	<i>Predictive targets are enriched for key functions associated with cardiotoxicity</i>	19
2.3.7	<i>Predictive targets are enriched for markers of skin and liver diseases</i>	19
2.4	<i>Discussion</i>	20
2.5	<i>Acknowledgements</i>	24
CHAPTER 3: KNOWLEDGE-GUIDED DEEP LEARNING MODELS OF DRUG TOXICITY IMPROVE INTERPRETATION		
3.1	<i>Introduction</i>	25
3.2	<i>Materials and Methods</i>	27
3.2.1	<i>Processing Tox21 datasets and inferring feature profile for DTox training</i>	27
3.2.2	<i>Constructing VNN with Reactome pathway hierarchy</i>	27
3.2.3	<i>Learning optimal DTox model for Tox21 assay outcome prediction</i>	29
3.2.4	<i>Interpreting optimal DTox model by layer-wise relevance propagation</i>	31
3.2.5	<i>Identifying significant VNN paths for explaining toxicity outcome of compounds</i>	32
3.2.6	<i>Comparing DTox against existing interpretation methods regarding rediscovering mechanisms of transcription activation by nuclear receptor</i>	33
3.2.7	<i>Processing LINCS dataset for validation of DTox interpretation results</i>	34
3.2.8	<i>Processing datasets for analyzing DTox results on HepG2- and HEK293-cytotoxic compounds</i>	35
3.3	<i>Results</i>	36
3.3.1	<i>Training DTox for predicting compound response to toxicity assays</i>	36
3.3.2	<i>DTox can achieve the same level of performance as complex classification algorithms</i>	41
3.3.3	<i>Development of a DTox interpretation framework for explaining VNN predictions</i>	44
3.3.4	<i>DTox can rediscover mechanisms of transcription activation by four nuclear receptors</i>	45

3.3.5	<i>DTox can recapitulate cellular activities induced by aromatase inhibitors and pregnane X receptor agonists</i>	48
3.3.6	<i>DTox can differentiate distinctive mechanisms leading to HepG2 cytotoxicity</i>	50
3.3.7	<i>Interpretation of HepG2 cytotoxicity links clinical phenotypes of DILI to TLR3/4 mediated necrosis</i>	52
3.3.8	<i>DTox can be applied to a wide range of chemicals other than drugs</i>	54
3.3.9	<i>HepG2 cytotoxicity scores predicted by DTox can differentiate hepatic cyst compounds from negative controls</i>	56
3.3.10	<i>DTox offers flexibility in balancing between model efficiency and performance</i>	57
3.4	<i>Discussion</i>	58
3.5	<i>Acknowledgements</i>	60
CHAPTER 4: KNOWLEDGE GRAPH AIDS COMPREHENSIVE EXPLANATION OF DRUG TOXICITY		61
4.1	<i>Introduction</i>	61
4.2	<i>Materials and Methods</i>	62
4.2.1	<i>Processing cell viability screening datasets for model training</i>	62
4.2.2	<i>Extracting chemical-gene connections from ComptoxAI for model construction</i>	62
4.2.3	<i>Constructing VNN with selected gene features and Reactome pathway hierarchy</i>	63
4.2.4	<i>Learning optimal AIDTox model for cytotoxicity prediction and interpretation</i>	64
4.3	<i>Results</i>	66
4.3.1	<i>AIDTox employs curated chemical-gene connections to construct VNN</i>	66
4.3.2	<i>Chemical-gene binding connections result in the best performing models</i>	67
4.3.3	<i>AIDTox models benefit from a comprehensive gene feature space</i>	69
4.3.4	<i>New features in AIDTox are essential in drug metabolism and elimination processes</i>	69
4.4	<i>Discussion</i>	70
4.5	<i>Acknowledgements</i>	72
CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS		73

5.1 Summary and highlights of TargetTox73
5.2 Limitations of TargetTox and future directions74
5.3 Summary and highlights of DTox76
5.4 Limitations of DTox and future directions77
5.5 Summary and highlights of AIDTox.....79
5.6 Graph neural network can improve the performance of toxicity prediction ..80
5.7 Concluding remarks84
BIBLIOGRAPHY86

LIST OF TABLES

<i>Table 2.1 Comparing model performance for 15 adverse events associated with tissue/organ damage</i>	<i>18</i>
<i>Table 2.2 Enrichment of disease markers/therapeutics among predictive targets</i>	<i>20</i>
<i>Table 3.1 Hyperparameter tuning of classification algorithms</i>	<i>31</i>
<i>Table 3.2 Summary of 15 Tox21 datasets used in the study.....</i>	<i>39</i>
<i>Table 3.3 Comparison of model efficiency among classification algorithms.....</i>	<i>57</i>

LIST OF ILLUSTRATIONS

<i>Figure 2.1 Identifying predictive features with TargetTox</i>	12
<i>Figure 2.2 Predicting binding outcomes for 569 targets</i>	13
<i>Figure 2.3 Predicting binding outcomes for 569 targets (extended analysis)</i>	14
<i>Figure 2.4 Comparing binding prediction across target function classes</i>	14
<i>Figure 2.5 Predicting toxicity outcomes for 815 adverse events</i>	15
<i>Figure 2.6 Correlation between width of performance confidence interval and ratio of positive to negative samples</i>	16
<i>Figure 2.7 Cluster map of 36 adverse events</i>	17
<i>Figure 2.8 Expression of predictive targets in the tissue of toxicity</i>	19
<i>Figure 2.9 Involvement of predictive targets in adverse outcome pathways</i>	22
<i>Figure 3.1 Modeling compound response to toxicity assay with DTox</i>	36
<i>Figure 3.2 Comparison of DTox model statistics</i>	37
<i>Figure 3.3 Evolution of loss function during learning of optimal DTox model</i>	38
<i>Figure 3.4 Prediction of compound response to 15 toxicity assays</i>	40
<i>Figure 3.5 Influence of pathway knowledge and hierarchy on predictive performance of DTox</i>	42
<i>Figure 3.6 Consistency of DTox interpretation across hyperparameter settings</i>	43
<i>Figure 3.7 Validation of identified VNN paths by known mechanisms</i>	45
<i>Figure 3.8 Validation of identified VNN paths by differential expression</i>	47
<i>Figure 3.9 In-depth analysis of HepG2 cytotoxicity using identified VNN paths</i>	49
<i>Figure 3.10 Clustering of HepG2-cytotoxic compounds based on cell death-related pathways</i>	50
<i>Figure 3.11 Summary of VNN paths identified for 413 cytotoxic compounds not mapped to cell death-related pathways</i>	51
<i>Figure 3.12 Application of predicted HepG2 cytotoxicity score among DSSTox compounds</i>	54
<i>Figure 3.13 Application of predicted HEK293 cytotoxicity score among DSSTox compounds</i>	55
<i>Figure 3.14 Influence of early stopping criterion on DTox model efficiency and performance</i>	57
<i>Figure 4.1 Incorporating curated chemical-gene connections into VNN for toxicity prediction with AIDTox</i>	65
<i>Figure 4.2 Relationship between training performance and the number of top predictive gene features</i>	66
<i>Figure 4.3 Comprehensive explanation of HEK293 cytotoxicity with new features in AIDTox</i>	68
<i>Figure 4.4 AIDTox explanation for HEK293 cytotoxicity of dasatinib</i>	71
<i>Figure 5.1 Overall performance metrics of the 3 QSAR model types on each of the Tox21 assays</i>	82

Figure 5.2 Receiver Operator Characteristic (ROC) curves for two selected Tox21 assays using different configurations of the GNN model.....83

CHAPTER 1: INTRODUCTION

1.1 Toxicity is a major cause of attrition in drug development

Drug toxicity refers to any untoward medical occurrence which does not necessarily have a causal relationship with the treatment. It is a primary reason for attrition in drug development, accounting for 30% of clinical trial failures in the last three decades¹⁻³. Postmarketing surveillance data showed that adverse drug events affect two million patients in the U.S. every year while causing approximately 100,000 fatalities⁴. Therefore, assessing toxicity is a critical step in drug development.

1.2 Toxicity can be caused by distinctive underlying cellular mechanisms

Based on the induced pathological effect, drug toxicity can be classified into four categories: cytotoxicity/tissue injury, altered phenotype/function, immunological hypersensitivity, and genotoxicity/cancer⁵. Drug toxicity also occurs in various contexts, including drug overdose, drug-drug interactions, rare idiosyncratic reactions representing unique susceptibility of individuals, and most commonly, adverse reactions at therapeutic doses. Toxicity can be caused by distinctive mechanisms. In general, there are four categories with regard to cellular mechanisms underlying toxicity: biological transformation to toxic metabolites, hypersensitivity and related immunological reactions, disruption of cellular signaling pathways related to on-target or off-target effects, and idiosyncratic reactions⁵.

Biological transformation (i.e. Phase I and II metabolism) is a metabolic process that xenobiotics undergo in human body. It is mediated by drug-metabolizing enzymes such as cytochrome P450 monooxygenases, aldehyde oxidase, UDP-glucuronosyltransferases, sulfotransferases, etc⁶. Reactive oxygen species generated from the process can cause oxidative stress by covalently binding to cell macromolecules such as DNA, proteins, and lipids⁷. This covalent binding may lead to cytotoxicity via perturbation of cellular signaling events, or genotoxicity via formation of DNA adducts.

Drug hypersensitivity is an immune-mediated reaction to a drug. It is often caused by covalent binding between drugs and serum or cell-bound proteins such as major

histocompatibility complex (MHC) class II molecules⁸. This covalent binding can activate T cells and trigger immune response such as antidrug antibody production, which results in symptoms ranging from mild skin rashes, anaphylaxis, to organ failure.

Drug-induced disruption of cellular signaling pathways can be a result of exaggerated pharmacological response at the therapeutic target (i.e. on-target effect). For instance, cardiac and pulmonary toxicity of tyrosine kinase inhibitors are linked to the disruption of EGFR signaling^{9, 10}. It can also be a result of interaction with unintended targets (i.e. off-target effect). For instance, hepatic and renal toxicity of antiretroviral agents are linked to the disruption of drug-metabolizing cytochrome P450 system^{11, 12}. Cardiac arrhythmia of non-sedating antihistamine terfenadine is linked to inhibition of cardiac ion channels (hERG)¹³.

Idiosyncratic reactions refer to adverse events that infrequently occur among individual patients or small subpopulations. The toxicities often reflect individual genetic predispositions that affect biological transformation, drug hypersensitivity, on-target or off-target effects^{14, 15}. This category of drug toxicity is the least well-understood since the reactions are rare (< 1 event in 10,000 individuals), and often not observed until late stage of clinical trials or post market introduction.

For most chemicals, the exact cellular mechanism of toxicity remains largely unknown¹⁶. Therefore, one key goal of toxicity assessment is to identify targets and pathways that may explain the outcome.

1.3 Large-scale toxicological data opens the door for computational research

There are three major data sources of large-scale toxicity assessment: preclinical studies, clinical trials, and postmarketing surveillance. Preclinical studies utilize quantitative high-throughput screening techniques to generate millions of data points regarding the response of biological systems to important chemical libraries, both *in vitro* and *in vivo*. For instance, in the Tox21 program¹⁷, over 8,500 compounds were tested for 72 toxicity endpoints including stress response, genotoxicity, cytotoxicity, developmental toxicity, etc. Similarly, in the ToxCast program¹⁸, 641 environmental chemicals and 135 reference pharmaceuticals were tested for 87

endpoints covering molecular functions related to cardiovascular disease, chronic inflammation, respiratory diseases, etc. These toxicity profiles can assist with probing how chemicals interact with proteins and pathways to trigger a certain outcome, and thus shed light on cellular mechanisms of toxicity¹⁹. Both clinical trials and postmarketing surveillance monitor the subjects' response to pharmaceuticals and keep track of the adverse events observed in subjects reflecting drug toxicity to human body systems, tissues, and cell types. Curated databases such as AACT²⁰, SIDER²¹, AEOLUS²², nSIDES²³, contain relationships between thousands of drugs and tens of thousands of adverse events. These resources provide extensive data for researches to model drug toxicity and develop *in silico* approaches identifying the chemical or biological patterns of a drug that might be predictive of adverse health outcomes in humans^{24, 25}.

1.4 Various *in silico* models have been developed for toxicity assessment

Existing strategies include *in vitro*, *in vivo*, and *in silico* testing. *In vitro* testing uses *in vitro* model systems (e.g. assays, organ-on-a-chip) to determine the potential of a chemical to be hazardous to humans. The development of high-throughput screening assays enables activity profiling of large numbers of chemicals simultaneously^{26, 27}. Recent advances in physiologically relevant 3D tissue models further enhance the translation of assay results to predict adverse effects in humans²⁸. *In vivo* testing uses model organisms (e.g. rat, mouse, rabbit) to determine the potential of a chemical to be hazardous to humans. *In vitro* and *in vivo* approaches are time-consuming, labor-intensive, and may not reflect the actual response in human body^{29, 30}. *In silico* testing uses computational models to perform large-scale virtual screening, identifying candidates for further experimental testing³¹.

A variety of *in silico* approaches have been developed for toxicity assessment, including rule-based structural alerts (SAs), uncertainty factors modeling (UFs), pharmacodynamics (PD) and pharmacokinetics (PK) modeling, Read-across, and ligand-based quantitative structure-activity relationship (QSAR) modeling³².

Rule-based SAs (i.e. toxicophores) contain structural properties that indicate certain toxicity outcome³³⁻³⁵. SAs are binary (the property is either present or absent) and only apply to

qualitative endpoints (e.g. cytotoxic or non-cytotoxicity)³⁴. They are not exclusive (absence of the property does not indicate non-toxicity), which may bring in many false negatives³³. Nor do they provide biological insights into the mechanism of toxicity.

UF models assess exposure risk or recommended intake of chemicals by interspecies extrapolation (from model organism to human), intraspecies extrapolation (from general population to particular groups such as elderly people, children, pregnant women), or exposure duration extrapolation (from short-term exposure to long-term exposure)^{36, 37}. UF models use a specified value (the uncertainty factor) to account interspecies/intraspecies/exposure duration variability. However, determination of UF is challenging as the factor itself varies greatly by toxicity endpoints. The adoption of empirical values may lead to inaccurate estimation.

PD models quantify drug effect on human body by relating the biological response to chemical concentration in tissue, whereas PK models quantify human body effect on drug (absorption, distribution, metabolism, and excretion processes) by relating chemical concentration in tissues to time³⁸⁻⁴⁰. Instead of administrated doses, PD and PK models take into account internal doses and key metabolites of a drug, which allow a more direct relationship with the drug response³⁹. However, PK and PD parameters are often estimated by *in vitro*-to-*in vivo* or interspecies extrapolation due to an absence of human data, assuming the dose-response relationships are consistent across species^{41, 42}. The assumption may lead to inaccurate estimation of concentrations in human body.

Read-across models use toxicity profiles of data-rich chemicals to infer the outcomes for data-poor chemicals⁴³⁻⁴⁵. Read-across is easy to implement and interpret. However, it largely depends on existing knowledge. The inference can be problematic when there is a lack of analog chemicals or the analogs have conflicting profiles⁴³.

The most commonly used *in silico* approach is ligand-based quantitative structure-activity relationship (QSAR) modeling⁴⁶. QSAR models quantify chemical structure of each drug into features, then relate the features to toxicity outcomes using supervised learning algorithms. The features can be binary chemical fingerprints indicting the yes/no answer to a set of questions

related to chemical structure (e.g. whether the chemical has more than three oxygens, whether the chemical has a disulfide bond), or continuous molecular descriptors representing structural properties of chemicals (e.g. molecular weight, partition coefficient). A wide range of supervised learning algorithms have been adopted to model the toxicity endpoints from structural features, including *k*-nearest neighbors^{47, 48}, Bayesian matrix factorization⁴⁹, support vector machines^{48, 50}, random forests^{47, 51}, gradient boosting⁵², and more recently, deep neural networks⁵³⁻⁵⁵.

1.5 Conventional QSAR models are limited by low accuracy and lack of interpretability

Many factors can affect the performance of QSAR models, including the chemical features used, the supervised learning algorithm employed, the composition of training set, and the endpoint of interest²⁵. QSAR models proved effective in predicting well-established *in vitro* toxicity endpoints⁵⁶, but often fell short of accuracy when used to predict complex *in vivo* endpoints such as drug adverse events⁵⁷. This can be attributed to the relatively low structure similarity among drugs causing the same adverse events⁵⁸, as well as the largely uncharacterized pharmacokinetics processes drugs undergo *in vivo*.

Regardless of the predictive performance, almost none of the existing QSAR models could overcome the trade-off between accuracy and interpretability. As algorithmic design gets more complex, it becomes challenging to interrogate how each input feature contributes to the eventual prediction⁵⁹. In addition, most interpretations only return a rank of structural properties or a set of classification rules⁵⁹; such interpretations are insufficient to explain the cellular mechanisms of toxicity. Integrated models have been proposed to combine structure properties with other feature types for toxicity prediction, including therapeutic targets⁴⁹, transcriptome response to drug-induced experiments⁵⁷, *in vitro* response to toxicity assays⁴⁸, dose-response relationships⁴⁷, etc. Though integrated models show improved performance over QSAR, they suffer from poor generalizability as prior knowledge is required for model construction.

In summary, most conventional *in silico* approaches for toxicity assessment are limited by low accuracy and lack of interpretability. Further, they often fail to explain cellular mechanisms

underlying structure-toxicity associations. Therefore, an accurate and interpretable model is urgently needed, which can connect drugs to their toxicity targets and pathways, generate new hypotheses for further testing, and facilitate mechanistic investigation.

CHAPTER 2: FEATURE SELECTION PIPELINE IDENTIFIES PREDICTIVE TARGETS ASSOCIATED WITH DRUG TOXICITY

This chapter was originally published as: Hao, Yun, and Moore, Jason, H. "*TargetTox: A Feature Selection Pipeline for Identifying Predictive Targets Associated with Drug Toxicity.*"

Journal of Chemical Information and Modeling. 2021 Nov 10;61(11):5386-94. doi:

10.1021/acs.jcim.1c00733

Contributions:

J.H.M. and Y.H. conceived the project. J.H.M. and Y.H. designed the study. Y.H. performed the analysis. J.H.M. and Y.H. interpreted the results and wrote the paper.

2.1 Introduction

Despite limited success in toxicity prediction, QSAR models showed promising performance in predicting compound-target interactions⁶⁰⁻⁶³. Target profile derived from QSAR models may indicate toxicity outcomes, since many pharmacovigilance studies have linked adverse drug events to aberrant activities of certain target proteins^{9-12, 64, 65}. Over the last decade, functional assays have generated extensive knowledge on compound-target interactions, enabling us to incorporate target knowledge into toxicity prediction. Nevertheless, the task remains challenging due to the contrast between high-dimensional target space and limited number of samples. To tackle this issue, we adopted ReBATE (Relief-based Algorithm Training Environment)⁶⁶ for implementing Relief-based feature-ranking methods on high-dimensional datasets. ReBATE methods rank features based on value difference between neighboring instances. Specifically, difference within same class contributes negatively to feature relevance while difference between classes contributes positively. ReBATE methods have two advantages over traditional (e.g. chi-square, ANOVA, mutual information) and tree-based methods, making it a proper choice for our task. First, they do not make assumptions regarding population distributions of features, thus can be applied to both continuous (e.g. structure properties) and binary (e.g. target profile) features. Second, they evaluate each feature in the context of remaining features, thus may preserve and detect prevalent interactions among features. A benchmark study showed that ReBATE methods

can prioritize two-way epistasis out of 1000 features, which all the other methods failed to achieve⁶⁷. In this study, we incorporated ReBATE methods into a pipeline, namely TargetTox, that identifies predictive features for a given dataset. We first implemented TargetTox on 569 compound-target interaction datasets, in order to identify structure properties predictive of binding outcomes, and to derive target profile of drugs. We then implemented TargetTox on 815 drug-adverse event datasets to identify targets predictive of toxicity outcomes. We further linked the predictive targets to adverse events by showing their differential expression in the tissue of toxicity, as well as their enrichment of toxicity-related functions and disease markers. We concluded with a discussion of potential applications for some predictive targets, which emerge as new markers of organ toxicity. Our code and data be accessed at <https://github.com/EpistasisLab/TTTox>. Our novel pipeline may benefit future studies of high-dimensional datasets.

2.2 Materials and Methods

2.2.1 Building compound-target interaction datasets

We obtained binding affinity of compound-target pairs from BindingDB⁶⁸. We converted binding affinity values into binary outcomes by first quartile of the distribution from all known drug-target pairs⁶⁹. We focused our study on targets with at least 50 paired compounds. To quantify structure properties of compounds, we used 246 molecular descriptors that cover most interesting chemical features for drug discovery⁷⁰.

2.2.2 Building drug-adverse event datasets

We obtained Proportional Reporting Ratio (PRR) of drug-adverse event pairs from OFFSIDES²³, along with the 95% confidence interval. PRR measures the extent to which an adverse event is disproportionately reported for individuals taking a given drug. We assigned drugs into case group (lower bound of PRR > 1) and control group (lower bound of PRR < 1, upper bound of PRR > 1). We focused our study on adverse with at least 500 paired drugs. To quantify structure properties of drugs, we used same set of 246 molecular descriptors as above.

2.2.3 Incorporating ReBATE methods to build a feature selection pipeline

The pipeline, namely TargetTox, starts by splitting each dataset into training and validation by ratio of four to one, then further splits the training set into ten folds. Rank of feature relevance is obtained by implementing ReBATE on data of nine folds. TargetTox then uses top-ranked features to fit classification models recursively, until an optimal performance (measured by testing AUROC on data of the remaining fold) can be reached. After repeating the above procedures for all ten folds, TargetTox selects predictive features that consistently appear across folds. The performance of selected features is evaluated on the held-out validation set.

2.2.4 Identifying optimal setting of hyperparameters by grid search

TargetTox has five hyperparameters. A brief description about each hyperparameter is given below:

- (i) Feature-ranking method. Two methods from ReBATE package were considered: MultiSURF and MultiSURFstar. MultiSURF only uses near instances to weigh feature relevance for a given target instance while MultiSURFstar takes into account both near and far instances.
- (ii) Whether to implement the “iterative scoring” function of ReBATE. Iterative scoring removes low-ranking features at each iteration, then reassigns neighbors based on the remaining features. The function is effective when a dataset contains a large number of features.
- (iii) Classification model. Two classification models were considered: random forest and gradient boosting.
- (iv) Tolerance score: maximal iterations to wait after last time testing AUROC improves, i.e. optimal performance is reached. Two values were considered: 20 and 50.
- (v) Consistency score: minimal proportion of folds in which a predictive feature appears. Two values were considered: 0.5 and 0.7.

Combined together, a total of 32 settings were considered for hyperparameter tuning. We randomly selected 100 compound-target interaction datasets for the task, then identified optimal setting by median training performance (measured by AUROC) across the 100 datasets. The identified optimal setting is (in the order listed above): MultiSURF, not to implement iterative scoring, random forest, tolerance score of 50, and consistency score of 0.5.

2.2.5 Implementing TargetTox to identify targets predictive for adverse events

With the optimal setting, we implemented TargetTox on compound-target interaction datasets to identify molecular descriptors predictive of binding outcomes. We fit classification models with the predictive descriptors, then implemented models with good validation performance (AUROC > 0.85) to derive target profile of all drugs. Finally, using the target profile as features, we implemented TargetTox on drug-adverse event datasets to identify targets predictive of toxicity outcomes. We computed 95% confidence interval for AUROC by generating bootstrapped samples from predicted outcome probabilities. On average, the bootstrapped samples contain 63.3% of unique original samples.

2.2.6 Comparing similarity of predictive descriptors between targets

We obtained class annotation of target proteins from dGene⁷¹ and GtoPDB⁷². We used Jaccard index to measure similarity of predictive descriptors between target pairs. We performed Mann-Whitney *U* test to examine whether target pairs of same class exhibit higher feature similarity than those of different classes. We obtained function annotation of target proteins from Gene Ontology (GO)^{73, 74} and Reactome⁷⁵. We removed function terms with either too few (< 10) or too many (> 100) annotated genes. Similarly, we performed Mann-Whitney *U* test to examine whether target pairs with common terms exhibit higher feature similarity than those without common terms. We then corrected for multiple comparisons by false discovery rate (FDR).

2.2.7 Comparing predictive targets identified by TargetTox to DisGeNET

We obtained 1,134,943 gene-disease associations from DisGeNET⁷⁶. The associations were integrated from four types of source databases: (i) expert curated resources (curated), (ii) resources derived from rat and mouse models of disease (animal), (iii) resources inferred from

human phenotype and genetic variant data (inferred), and (iv) resources extracted from previous literature using text mining tools (literature). After matching disease names with adverse events, we performed Fisher's exact test to examine the significance of overlaps between disease-associated genes and predictive targets of each adverse event. We then corrected for multiple comparisons by FDR.

2.2.8 Clustering adverse events by predictive targets

We performed average linkage hierarchical clustering on adverse events with good validation performance (AUROC > 0.65). We measured distances between adverse events by Jaccard distance in target space, which comprises predictive targets appearing in at least five adverse events. We implemented R package Pvcust⁷⁷ to identify clusters from the dendrogram of hierarchical clustering. Pvcust computes p-value for each cluster using bootstrap resampling techniques. To account for the stochastic nature of resampling, we ran the analysis for 20 times and identified clusters that repeatedly appear as significant ($P < 0.05$). We focused our analysis on small clusters (< 10 adverse events).

2.2.9 Analyzing differential expression of predictive targets in tissue of toxicity

We obtained mRNA expression (measured by Transcripts Per Million) data of human tissues from GTEx⁷⁸. We removed genes with zero expression in all tissues, then adjusted for baseline expression of each remaining gene by:

$$e_{tissue-adj} = \left| \log_{10} \frac{e_{tissue}}{e_{median}} \right|$$

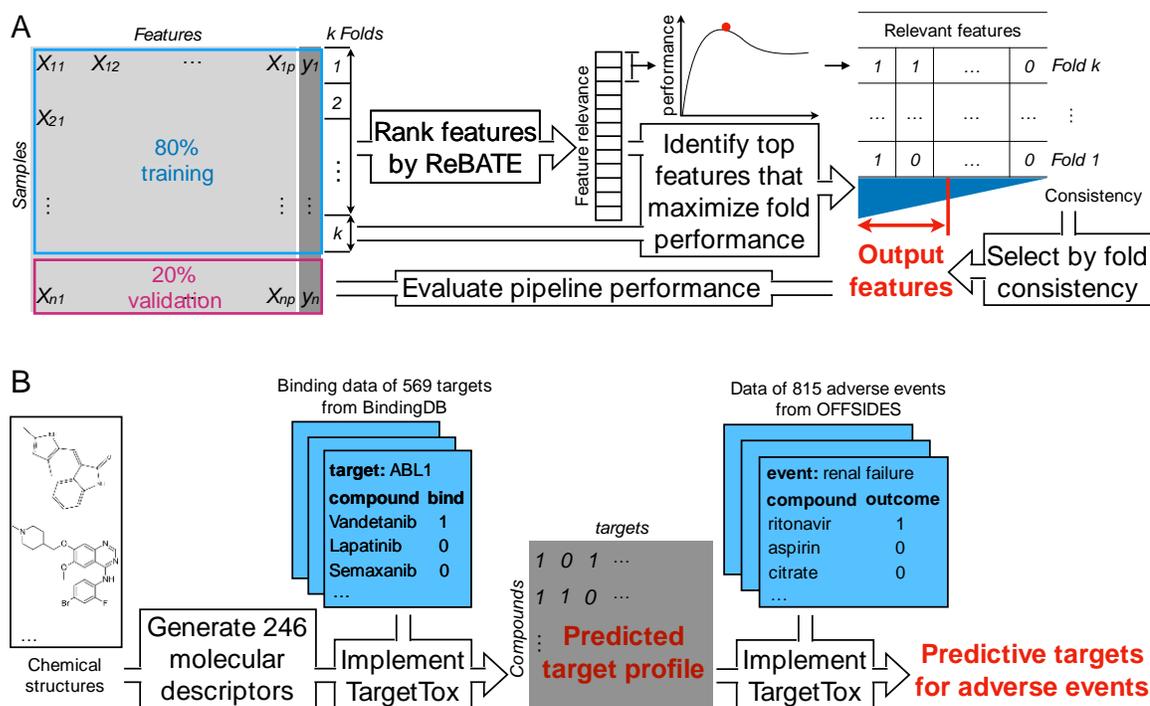
where e_{tissue} and e_{median} are the expression in a given tissue and the median expression across all tissues, respectively. The adjusted value measures the extent to which a gene is differentially expressed in a given tissue. We mapped each adverse event to its tissue of toxicity, then performed Mann-Whitney U test to examine whether predictive targets exhibit higher expression than background genes. We then corrected for multiple comparisons by FDR.

2.2.10 Identifying enriched GO terms for predictive targets

We performed GO enrichment analysis on predictive targets of each adverse event by Fisher's exact test. We focused our analysis on GO terms with size between 10 and 100. We conducted the analysis separately for three GO branches: biological process, molecular function, and cellular component. We then corrected for multiple comparisons by FDR.

2.2.11 Identifying disease genes from predictive targets

We obtained gene-disease connections from CTD⁷⁹. We focused our analysis on connections with direct evidence (i.e. the gene is a disease marker or therapeutic). We mapped each adverse event to its disease category, then used key words to search for disease terms of each category along with their related genes. We performed Fisher's exact test to examine overrepresentation of disease-related genes among predictive targets of each adverse event. We then corrected for multiple comparisons by FDR.



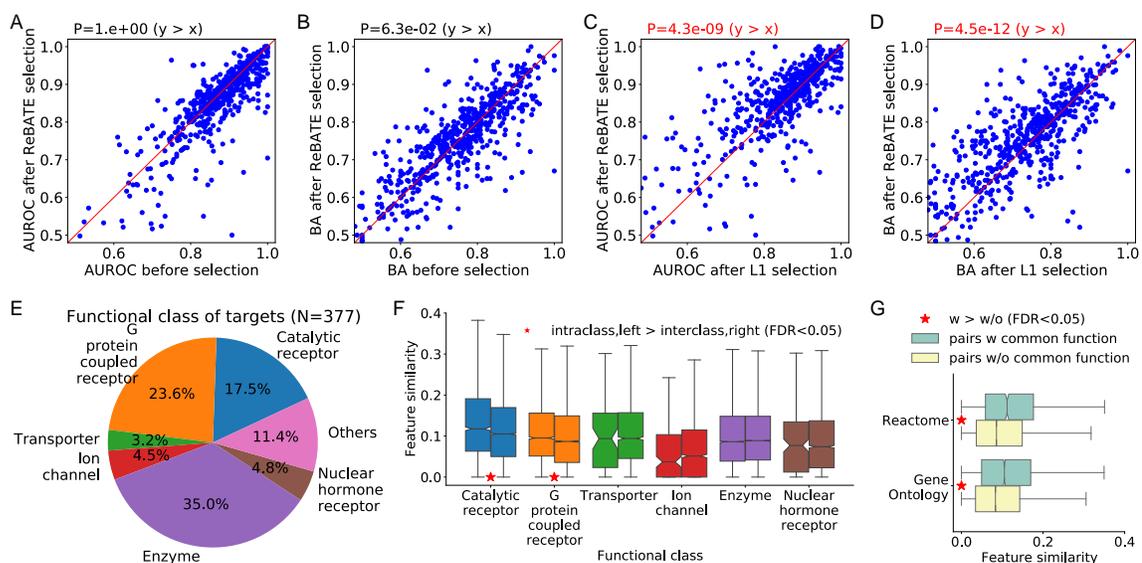


Figure 2.2 Predicting binding outcomes for 569 targets

(A-D) Scatter plots comparing model performance. Each point represents the comparison for one target. Prediction was evaluated in two performance metrics: AUROC (A&C) and balanced accuracy (B&D). TargetTox was compared against two other models: i) a model with same procedure as TargetTox except for no feature selection (A&B), and ii) a model with same procedure as TargetTox except for substituting ReBATE with L1 regularization (C&D). Wilcoxon signed-rank test was employed to examine whether TargetTox (y-axis) outperformed the compared model (x-axis), with p-value shown on the top. Significant comparison ($P < 0.05$) was highlighted in red. **(E)** Pie chart showing function class distribution among 377 targets. These targets have prediction models with AUROC > 0.85. **(F)** Boxplot comparing feature similarity within and between classes. Feature similarity was measured by Jaccard index of predictive descriptors between target pairs (y-axis), then compared across six classes (x-axis). The notches represent 95% confidence interval around the median. Mann-Whitney U test was employed to examine whether target pairs within each class (left box) exhibit higher feature similarity than those between classes (right box). Significant comparison ($FDR < 0.05$) was highlighted with a red star. **(G)** Similar to (F). Comparison was made between target pairs with common annotations (green box) and those without common annotations (yellow box). Two annotation sources (y-axis) were considered: Gene Ontology and Reactome.

2.3 Results

2.3.1 TargetTox can accurately predict binding outcomes for targets

We implemented TargetTox (Figure 2.1A) to identify molecular descriptors that are predictive of binding outcomes for 569 targets (Figure 2.1B). TargetTox identified an average of 26 ± 2 predictive descriptors per target, accounting for ten percent of all descriptors. When used for binding outcome prediction, the predictive descriptors achieved an average AUROC of 0.86 ± 0.01 and an average balanced accuracy of 0.75 ± 0.01 , approximating the performance by all descriptors (Figure 2.2A&B). We compared TargetTox with two L1 regularization-based models: i) a random forest model built upon L1-selected features (Figure 2.2C&D), and ii) a logistic regression model built upon L1-selected features (Figure 2.3A&B). Note that the first model has

the same procedure as TargetTox except for substituting ReBATE with L1 regularization. As a result, TargetTox outperformed both models by a large margin (i: $P = 4.3e^{-9}$; ii: $P = 3.0e^{-46}$, Wilcoxon signed-rank test) with much fewer descriptors ($P = 3.4e^{-54}$, Wilcoxon signed-rank test; Figure 2.3C).

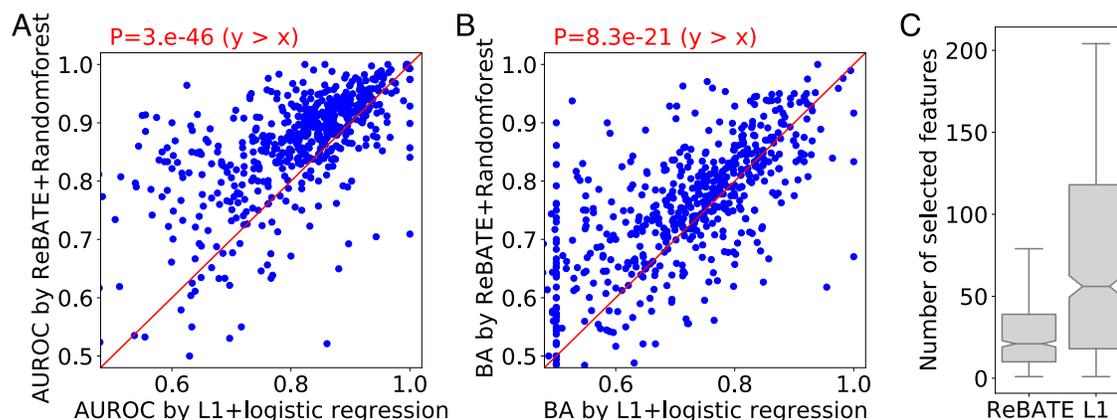


Figure 2.3 Predicting binding outcomes for 569 targets (extended analysis)

(A, B) Scatter plots comparing model performance. Each point represents the comparison for one target. Prediction was evaluated in two performance metrics: AUROC (A) and balanced accuracy (B). Our pipeline was compared to a logistic regression model built upon L1-selected features. Wilcoxon signed-rank test was employed to examine whether our pipeline (y-axis) outperformed the compared model (x-axis), with p -value shown on the top. (C) Boxplot comparing numbers of selected features. The comparison was made between two methods: (i) our ReBATE-based pipeline and (ii) L1-regularization.

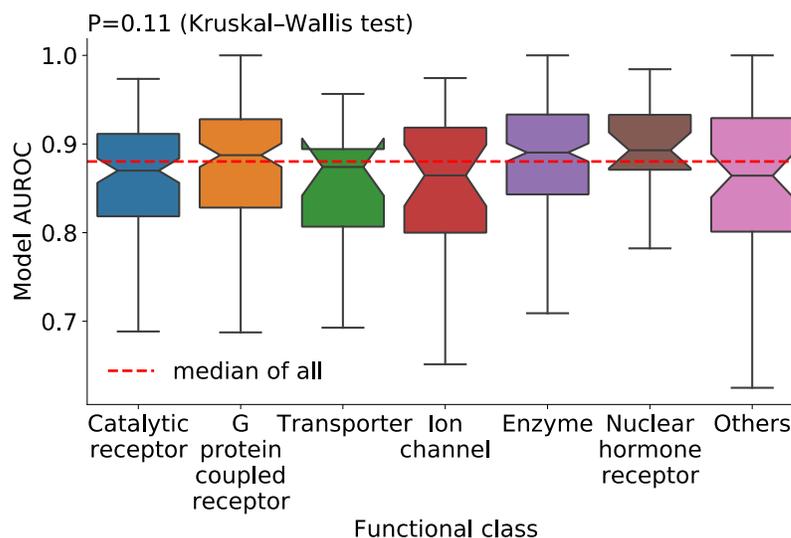


Figure 2.4 Comparing binding prediction across target function classes

Prediction was evaluated in AUROC (y-axis), then compared across seven function classes (x-axis). The notches represent 95% confidence interval around the median. Kruskal-Wallis test was employed to examine

whether distributions of AUROC vary by function class, with p-value shown on the top. The median performance across all classes is highlighted in a red dashed line.

2.3.2 TargetTox can identify similar structure properties for target proteins of similar function

Under TargetTox, 377 of 569 (66.3%) targets had binding prediction models with AUROC greater than 0.85 (Figure 2.2E). These target proteins expand across a variety of function classes: enzyme (35%), G-protein coupled receptor (23.6%), catalytic receptor (17.5%), nuclear hormone receptor (4.8%), ion channel (4.5%), and transporter (3.2%). The performance of TargetTox does not vary by function class ($P = 0.11$, Kruskal-Wallis test; Figure 2.4).

We compared similarity of predictive descriptors within and between function classes (Figure 2.2F). The comparison showed that target pairs within two function classes (G-protein coupled receptor and catalytic receptor) are more likely to share predictive descriptors (FDR < 0.05, Mann-Whitney U test). The two classes make up 41 percent of all targets being studied. We also compared similarity of predictive descriptors within and between function annotations (Figure 2.2G). Similarly, target pairs with common function annotations are more likely to share predictive descriptors (FDR < 0.05, Mann-Whitney U test).

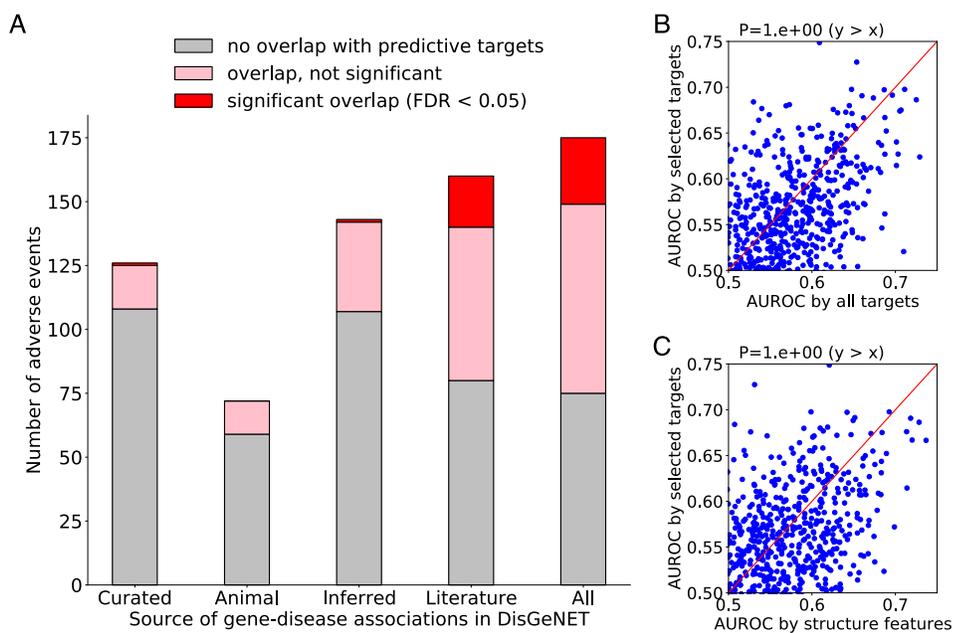


Figure 2.5 Predicting toxicity outcomes for 815 adverse events

(A) Barplot showing the number of adverse events in which the identified predictive targets overlap with disease-associated genes from DisGeNET. Gene-disease associations were collected from five types of resources (x-axis). Fisher's exact test was employed to examine the significance of overlap. (B-C) Scatter plots comparing model performance. Each point represents the comparison for one adverse event. Prediction was evaluated in AUROC. TargetTox was compared against two other models: i) a model with same procedure as TargetTox except for no feature selection (B), and ii) a model built upon molecular descriptors (C). Wilcoxon signed-rank test was employed to examine whether TargetTox (y-axis) outperformed the compared model (x-axis), with p-value shown on the top.

2.3.3 TargetTox can achieve same level of performance as QSAR in toxicity outcome prediction

We then implemented TargetTox to identify targets that are predictive of toxicity outcomes for 815 adverse events (Figure 2.1B). TargetTox identified an average of 26 ± 1 predictive targets per adverse event, accounting for seven percent of all target features. In DisGeNET, 175 of 815 adverse events were annotated with at least one disease-associated gene. We found significant overlaps between identified predictive targets and annotated genes in 26 (14.9%) adverse events, and non-significant overlaps in 74 (42.3%) adverse events (Figure 2.5A). We also analyzed the overlaps with disease-associated genes by annotation source. Overlaps were detected in fewer adverse events when considering expert-curated (18 of 126) or animal model-derived associations (13 of 72), but detected in more adverse events when considering phenotype-inferred (36 of 143) or literature-extracted (80 of 160) associations. In 75 (42.9%) adverse events, the identified predictive targets cannot be found in DisGeNET, suggesting new discoveries of associations by TargetTox.

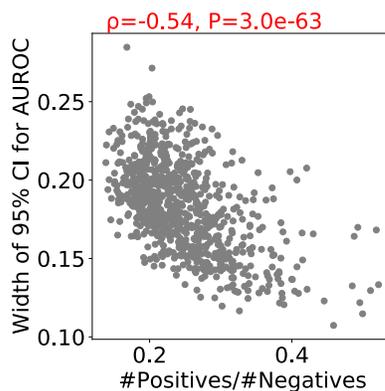


Figure 2.6 Correlation between width of performance confidence interval and ratio of positive to negative samples

The width of 95% confidence interval for AUROC (y-axis) was plotted against the ratio of positive to negative samples (x-axis). Each point represents the comparison for one adverse event. Spearman's ρ was employed to examine correlation between two axes, with p-value shown on the top.

When used for toxicity outcome prediction, the predictive targets achieved an average AUROC of 0.539 ± 0.005 . The average width of 95% confidence interval for AUROC is 0.182 ± 0.002 . The width is negatively correlated with the ratio of positive to negative samples (Spearman's $\rho = -0.54$, $P = 3.0e^{-63}$; Figure 2.6). We compared TargetTox with prediction models built upon all target features (Figure 2.5B). The comparison showed that feature selection by ReBATE did not improve the overall performance in toxicity outcome prediction ($P > 0.99$, Wilcoxon signed-rank test). We also compared TargetTox with a QSAR model built upon molecular descriptors (Figure 2.5C). Similarly, TargetTox did not outperform the QSAR model ($P > 0.99$, Wilcoxon signed-rank test).

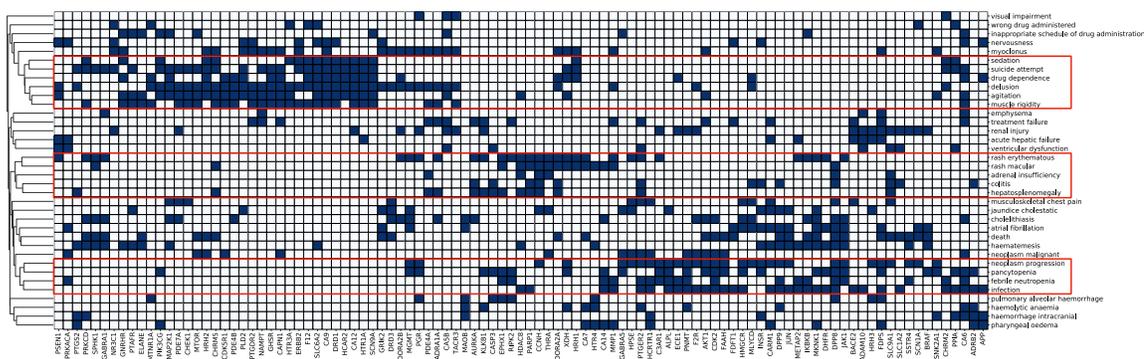


Figure 2.7 Cluster map of 36 adverse events

These adverse events (rows) have prediction models with AUROC > 0.65 . They are clustered by predictive targets (columns). Average linkage hierarchical clustering was employed to generate the map, with dendrogram shown on the left. Significant cluster detected by 'pvclust' was highlighted with a red box.

2.3.4 Similar adverse events can be clustered together by predictive targets

Under TargetTox, 36 of 815 (4.4%) adverse events had a prediction model with AUROC greater than 0.65. These adverse events expand across a variety of categories. We performed hierarchical clustering on the adverse events using a subset of predictive targets (Figure 2.7). Three significant clusters were identified. The first cluster (top) expands across six behavioural events: sedation, suicide attempt, drug dependence, delusion, agitation, and muscle rigidity. The second cluster (middle) expands across five events, including two dermal events: rash erythematous and rash macular, as well as two digestive events: colitis and hepatosplenomegaly.

The third cluster (bottom) expands across four events, including three hematologic events: pancytopenia, febrile neutropenia, and infection.

adverse event term	Model AUROC with 95% CI			#predictive targets
	246 molecular descriptors	377 target features (before selection)	predictive targets (after selection)	
muscle rigidity	0.69±0.10	0.71±0.09	0.70±0.08	37
hepatosplenomegaly	0.64±0.11	0.69±0.12	0.70±0.10	18
atrial fibrillation	0.51±0.07	0.53±0.07	0.68±0.07	27
emphysema	0.61±0.10	0.54±0.10	0.68±0.08	10
acute hepatic failure	0.53±0.10	0.63±0.09	0.68±0.10	37
colitis	0.65±0.09	0.64±0.09	0.67±0.09	24
pancytopenia	0.63±0.07	0.63±0.07	0.67±0.06	48
musculoskeletal chest pain	0.55±0.09	0.49±0.10	0.67±0.09	22
rash erythematous	0.60±0.08	0.65±0.08	0.67±0.08	57
adrenal insufficiency	0.72±0.10	0.69±0.11	0.67±0.10	21
pulmonary alveolar haemorrhage	0.59±0.11	0.54±0.11	0.66±0.10	17
rash macular	0.52±0.11	0.58±0.10	0.66±0.09	26
renal injury	0.60±0.11	0.60±0.11	0.66±0.10	48
myoclonus	0.67±0.09	0.65±0.10	0.65±0.08	50
ventricular dysfunction	0.52±0.10	0.53±0.10	0.65±0.11	13

Table 2.1 Comparing model performance for 15 adverse events associated with tissue/organ damage

2.3.5 Predictive targets are differentially expressed in the tissue of toxicity

Among the 36 adverse events mentioned above, 15 of them are associated with damage in a specific tissue/organ (Table 2.1). We studied the mRNA expression of predictive targets in the tissue of toxicity (Figure 2.8). For seven of the 15 adverse events, predictive targets exhibit differential expression (FDR < 0.05, Mann-Whitney *U* test) in the matched tissue: colitis (colon), renal injury (kidney), muscle rigidity (muscle), myoclonus (muscle), rash erythematous (skin), rash macular (skin), and pancytopenia (blood).

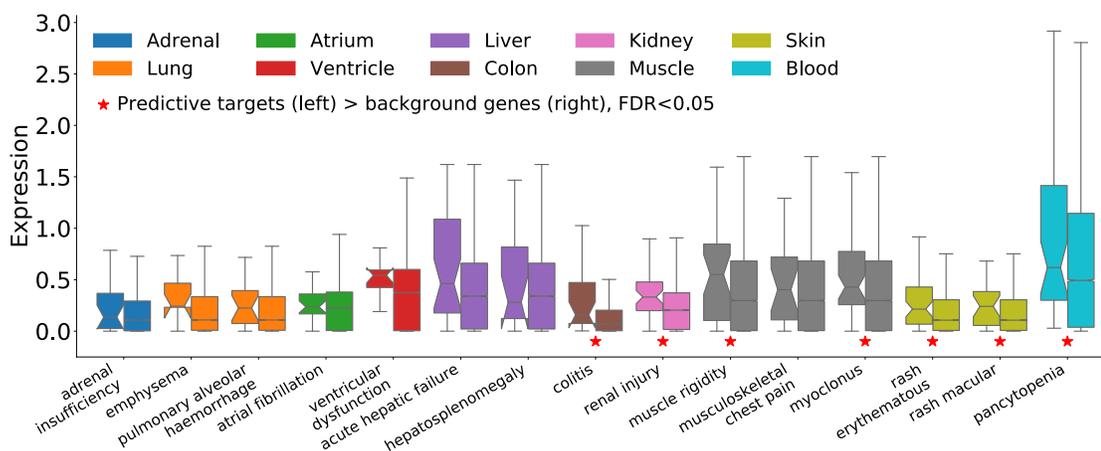


Figure 2.8 Expression of predictive targets in the tissue of toxicity

Tissue-specific expression (y-axis) was compared between predictive targets and background genes for 15 adverse events (x-axis). Each adverse event was mapped to its tissue of toxicity. Boxes were colored by tissue, with legend shown on the top. The notches represent 95% confidence interval around the median. Mann-Whitney U test was employed to examine whether predictive targets (left box) exhibit higher expression than background genes (right box). Significant comparison (FDR < 0.05) was highlighted with a red star.

2.3.6 Predictive targets are enriched for key functions associated with cardiotoxicity

To understand tissue-specific functions of predictive targets, we performed GO enrichment analysis. The analysis revealed some key functions of predictive targets in the tissue of toxicity. For instance, targets predictive of atrial fibrillation are enriched for two GO terms (FDR = 0.04, Fisher's exact test), both of which are related to cholesterol biosynthesis ('regulation of cholesterol biosynthetic process' and 'cholesterol biosynthetic process')⁸⁰. Meanwhile, targets predictive of ventricular dysfunction are enriched for five GO terms (FDR < 0.05, Fisher's exact test), including a term related to ventricular development ('Notch receptor processing, ligand-dependent')⁸¹ and a term related to ventricular fibrillation ('cellular response to epinephrine stimulus')⁸².

2.3.7 Predictive targets are enriched for markers of skin and liver diseases

To further connect predictive targets to adverse events, we examined whether they are enriched for known markers/therapeutics of the matched disease category (Table 2.2). Among the 13 adverse events being studied, 11 of them contain at least one marker or therapeutic of the

matched disease category. Two adverse events, namely rash erythematous and hepatosplenomegaly, are significantly enriched for markers/therapeutics of skin and liver diseases, respectively (FDR < 0.05, Fisher's exact test). Three more adverse events, including colitis, ventricular dysfunction, and rash macular, are reported with $P < 0.05$ before adjustment. They are also more than twice likely to contain disease markers/therapeutics (Odds ratio > 2).

disease category	adverse event term	#predictive targets are marker	#predictive targets are not marker	#other targets are marker	#other targets are not marker	odds ratio	P value	FDR
skin	rash erythematous	8	42	337	8292	4.69	0.001	0.009
liver	hepatosplenomegaly	9	7	2048	6615	4.15	0.005	0.03
colon	colitis	3	19	241	8416	5.51	0.02	0.1
heart	ventricular dysfunction	4	8	1035	7632	3.69	0.05	0.13
skin	rash macular	3	18	342	8316	4.05	0.05	0.13
muscle	musculoskeletal chest pain	1	16	202	8460	2.62	0.33	0.62
muscle	myoclonus	2	45	201	8431	1.86	0.30	0.62
lung	pulmonary alveolar haemorrhage	2	13	862	7802	1.39	0.45	0.73
heart	atrial fibrillation	3	23	1036	7617	0.96	0.62	0.89
kidney	renal injury	2	40	517	8120	0.79	0.73	0.94
lung	emphysema	0	9	864	7806	0	1	1
liver	acute hepatic failure	4	30	2053	6592	0.43	0.98	1
muscle	muscle rigidity	0	34	203	8442	0	1	1

Table 2.2 Enrichment of disease markers/therapeutics among predictive targets

2.4 Discussion

Ligand-based QSAR model has two major limitations in drug adverse event prediction. First, it cannot provide accurate predictions for complex events due to a lack of high-quality training datasets, as well as the complex transformations of drugs in human body. Second, it cannot explain mechanisms of toxicity as structure properties do not always shed light on cellular activity of drugs. In contrast, target-based prediction appears to be a better alternative as drugs causing same adverse events often share targets. Therefore, we proposed a prediction model that relates

structure properties to toxicity outcomes via target profile of drugs. To resolve data dimensionality issues, we developed TargetTox that incorporates feature-ranking methods from ReBATE. We implemented TargetTox to identify predictive descriptors for target binding, as well as predictive targets for adverse events. In both tasks, we obtained similar sets of predictive features for outcomes of alike class/category. In 100 adverse events, we rediscovered at least one disease-associated gene from the identified predictive targets. With the predictive features, we were able to approximate the performance by all features in both binding and toxicity prediction. The predictive features make up less than 10 percent of original features, which makes our prediction model less vulnerable to overfitting and more interpretable. We also demonstrated the advantage of ReBATE methods over L1 regularization, as they resulted in better predictive performance with fewer features. However, we noticed that our target-based model did not significantly outperform QSAR in toxicity prediction. The relatively low AUROC is in line with previous efforts predicting drug adverse events^{57, 83}. In addition to metabolic transformations, whether a drug activates or inhibits its target *in vivo* can also play a critical role leading to some adverse events. Therefore, the binary target profile derived by TargetTox may have limited predictive power. While directed target profile cannot be generated with existing binding data, we expect the issue to be resolved when more target activation/inhibition data becomes available. Furthermore, the uncertainty of our predictions remains high as we observed large confidence intervals for AUROC. This can be attributed to the imbalanced ratio of positive to negative samples. Despite these limitations, overall, our target-based model was able to achieve the same level of performance as QSAR.

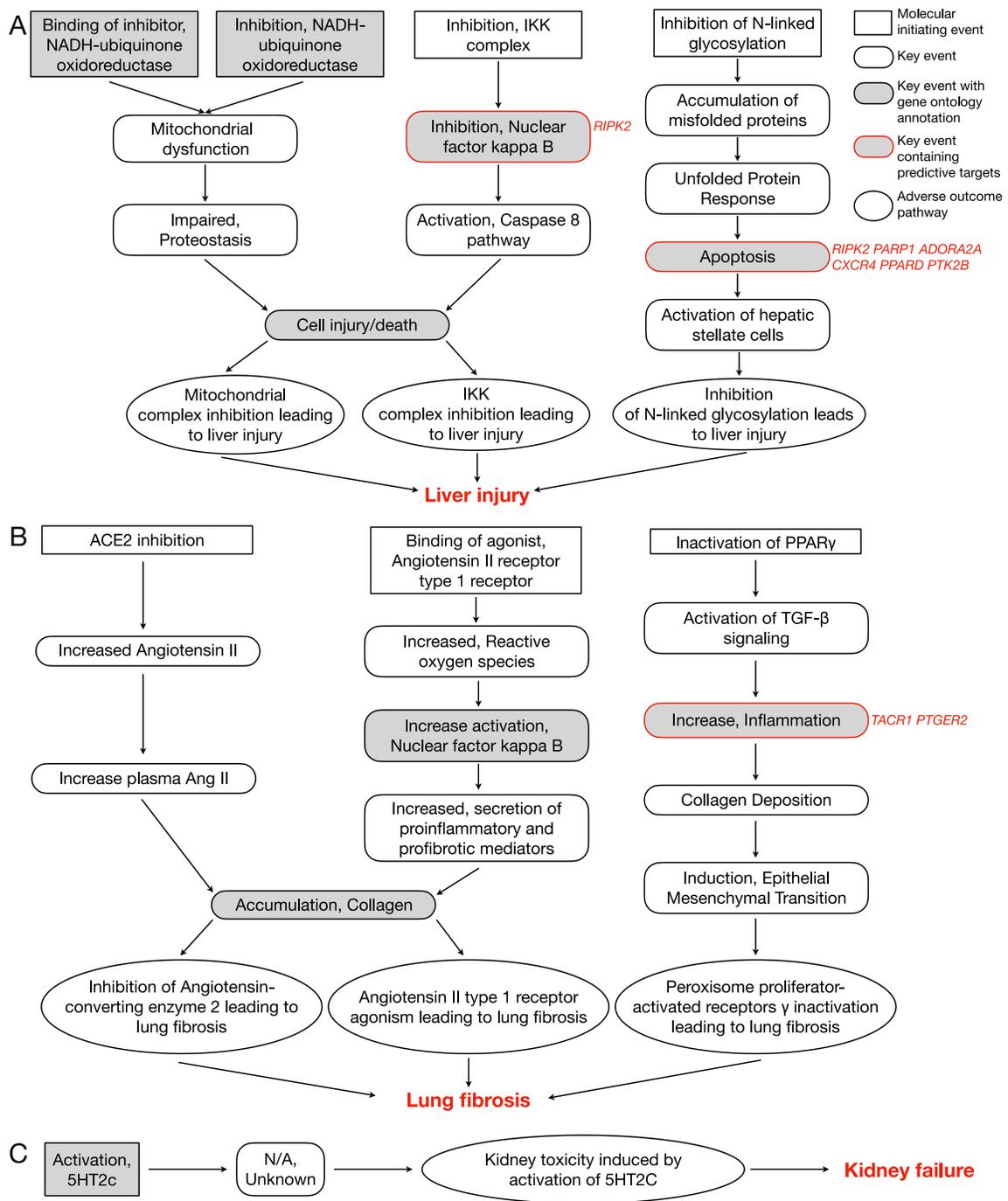


Figure 2.9 Involvement of predictive targets in adverse outcome pathways
 Adverse outcome pathways (AOPs) of three adverse events were shown in the graph (A: Liver injury, B: Lung fibrosis, C: Kidney failure). Each AOP is arranged in the order of molecular initiating event, key event, and outcome. Predictive targets involved in key events are shown next to the respective events.

While accuracy remains the same, interpretability has been significantly improved under our model. This is because our model is built upon a few predictive targets which can link chemicals to their cellular activities. To understand functions of the predictive targets, we studied their tissue-specific expression, functional annotations, and connections with known disease genes. We discovered that predictive targets tend to be differentially expressed in the tissue of toxicity, suggesting tissue-specific roles of predictive targets. We recapitulated some of the key roles in cardiotoxicity. For instance, we found targets predictive of ventricular dysfunction participate in Notch signaling and epinephrine signal transduction. Previous studies have indicated that Notch signaling is required for ventricular development to sustain cardiomyocyte proliferation⁸¹ while epinephrine is effective in treating prolonged ventricular fibrillation⁸². We demonstrated the predictive targets, especially those of rash and hepatosplenomegaly, are likely to be markers/therapeutics of the matched disease category. We also analyzed predictive targets in the context of adverse outcome pathways (AOPs) using data from AOPwiki⁸⁴. We found presence of six predictive targets in two of three established AOPs for liver injury: I κ B kinase complex inhibition⁸⁵ and N-linked glycosylation inhibition⁸⁶ (Figure 2.9A). Notably, we found RIPK2, connected to liver injury by TargetTox, participates in two key events: NF- κ B inhibition (leading to I κ B kinase complex inhibition) and apoptosis (leading to N-linked glycosylation inhibition). Similarly, we found presence of two predictive targets, namely TACR1 and PTGER1, in an established AOP for lung fibrosis (Figure 2.9B). The two targets can cause an increase in inflammation, which eventually leads to the inactivation of PPAR- γ , a suppressor for lung fibrosis⁸⁷. Combined together, these results imply that our predictive targets can help explain the cellular mechanisms of toxicity.

In addition to rediscovering known markers, TargetTox can also identify potential new markers of organ toxicity, as we found *in vitro* or *in vivo* evidence supporting the key roles of some predictive targets in liver, kidney, colon, and lung. For instance, we identified KLKB1, a plasma kallikrein with high expression in liver, to be predictive of hepatosplenomegaly. A recent study discovered that inhibition of KLKB1 reduced lipid deposition in hepatocytes, a risk factor for

hepatomegaly⁸⁸. We identified HSD11B1, a hydroxysteroid dehydrogenase with high expression in liver, to be predictive of acute hepatic injury. HSD11B1 limits activation of hepatic stellate cell, a major cell type involved in liver fibrosis. HSD11B1 inhibitors were found to promote fibrosis in murine liver injury⁸⁹. We identified CA12, a carbonic anhydrase with high mRNA expression in kidney, to be predictive of renal injury. Carbonic anhydrases participate in nitrite resorption while their inhibitors exhibited reno-protective effects in rat models of renal failure⁹⁰. We identified NPY2R, a Neuropeptide Y receptor that mediates nutrient absorption in the colon, to be predictive of colitis. It has been revealed that targeted deletion of Neuropeptide Y can modulate experimental colitis in mice⁹¹. We identified SLC5A1, a sodium/glucose cotransporter that modulates airway surface glucose concentration, to be predictive of pulmonary alveolar haemorrhage. Increased activity of SLC5A1 in lung alveolar cells has been shown to prevent pulmonary infection, a primary cause of alveolar haemorrhage⁹². These new discoveries highlight the potential of TargetTox for identifying targets with diagnostic/therapeutic applications. In the future, we believe TargetTox can be extended to study high-dimensional datasets from other domains.

2.5 Acknowledgements

This work was supported by NIH grant P30ES013508.

CHAPTER 3: KNOWLEDGE-GUIDED DEEP LEARNING MODELS OF DRUG TOXICITY IMPROVE INTERPRETATION

This chapter was originally published as: Hao, Yun, Romano, Joseph, D., and Moore, Jason, H., “*Knowledge-guided deep learning models of drug toxicity improve impretertation.*” *Patterns*. 2022 Sep 9; 3(9):100565. doi: 10.1016/j.patter.2022.100565

Contributions:

J.H.M. and Y.H. conceived the DTox project. J.H.M. and Y.H. designed the DTox model and data analysis workflow. Y.H. and J.D.R. performed the analysis (J.D.R. helped with the cytotoxicity analysis of DSSTox chemicals in 3.3.8 and 3.3.9). J.H.M. and Y.H. interpreted the results and wrote the paper with editing by J.D.R.

3.1 Introduction

Previous studies have modeled drug toxicity from physiochemical properties of compounds using a wide range of supervised learning algorithms, including k -nearest neighbors^{47, 48}, Bayesian matrix factorization⁴⁹, support vector machines^{48, 50}, random forests^{47, 51}, gradient boosting⁵², and more recently, deep neural networks⁵³⁻⁵⁵. Even though most of these algorithms achieved decent predictive performance, none of them could overcome the trade-off between accuracy and interpretability. As algorithmic design gets more complex, it becomes challenging to interrogate how each input feature contributes to the eventual prediction⁵⁹. A few post-hoc explanation techniques, such as local interpretable model-agnostic explanations (LIME)⁹³ and deep learning important features (DeepLIFT)⁹⁴, were developed to address the challenge. Nevertheless, these techniques often draw criticism in that they only provide an approximate explanation with locally fitted naïve models. Thus, they may not reflect the real behavior of original model⁹⁵. More critically, the setting of existing toxicity prediction models has limited the explanation of contributions from structural properties or target proteins while interactions with pathways remain largely uncharacterized. For toxicologists, the behavior of pathways proves crucial in deciphering the cellular activities induced by a compound, and understanding how target proteins, specific pathways, and biological processes trigger the toxicity outcome as a

whole²⁵. Therefore, a toxicity prediction model that achieves interpretability at both the gene and pathway level is urgently needed.

Recent developments in visible neural networks (VNN) have overcome the accuracy-interpretability trade-off. VNN is a type of neural network whose structure is guided by extensive knowledge from biological ontologies and pathways. The incorporation of ontological hierarchy in VNN forms a meaningful network structure that connects input gene features to output response via hidden pathway modules, making the model highly interpretable at both gene and pathway level. In a pioneering study, Ma *et al.* built a VNN with 2,526 Gene Ontology and Clique-eXtracted Ontology terms, for predicting growth rate of yeast cells from gene deletion genotypes⁹⁶. The authors were also able to rediscover key ontology terms responsible for cell growth by examining the structure of the VNN. Subsequent studies have extended the VNN model for learning tasks regarding human cells, such as predicting drug response and synergy in cancer cell lines⁹⁷, modeling cancer dependencies⁹⁸, and stratifying prostate cancer patients by treatment-resistance state⁹⁹. It is our working hypothesis that VNNs can address the limitations of existing toxicity prediction models due to their incorporation of pathway knowledge and the resulting high interpretability. In this study, we employed the Reactome⁷⁵ pathway hierarchy to develop a VNN model—namely DTox—for predicting compound response to 15 toxicity assays. Further, we developed a DTox interpretation framework for identifying VNN paths that can explain the toxicity outcome of compounds. We connected the identified VNN paths to cellular mechanisms of toxicity by showing their involvement in the target pathway of respective assay, their differential expression in the matched Library of Integrated Network-Based Cellular Signatures (LINCS) experiment¹⁰⁰, and their compliance with screening results from mechanism of action assays. We applied the DTox models of cell viability to perform a virtual screening of ~700,000 compounds and linked predicted cytotoxicity scores with clinical phenotypes of drug-induced liver injury. We conclude with a discussion of potential discoveries made by DTox, some of which have already been validated in previous studies. Our code can be accessed openly at <https://github.com/yhao->

compbio/DTox. In general, the DTox interpretation framework will benefit *in silico* mechanistic studies and generate testable hypotheses for further investigation.

3.2 Materials and Methods

3.2.1 Processing Tox21 datasets and inferring feature profile for DTox training

The Tox21 datasets¹⁷ contain screening results describing the response of *in vitro* toxicity assays to compounds of interest, including approved drugs, experimental drugs, small molecules, and environmental chemicals. We extracted active and inactive compounds from the screening results of each assay, then removed compounds with inconclusive or ambiguous results. We further removed assays with fewer than 5,000 available compounds, focused our analyses on the remaining 15 assays. To quantify structural properties of compounds, we used *rdck* package to compute a 166-bit binary MACCS fingerprint that covers most of the interesting physicochemical features for drug discovery¹⁰¹. We then implemented TargetTox¹⁰² to infer the target-binding probability of each compound from its MACCS fingerprint. TargetTox comprises binding prediction models that were pre-trained on hundreds of thousands of compound-target binding affinity data points that were experimentally measured in EC50/IC50/K_d/K_i. It first employs a feature selection pipeline to identify the fingerprint features that are predictive of the binding outcome for each target protein, then fits a random forest classification model using the predictive features. We selected 361 target proteins of which the binding outcome can be well predicted by TargetTox (model AUROC > 0.85 on held-out validation set). The derived target-binding profile containing 361 proteins were then used as input feature data for assay outcome modeling.

3.2.2 Constructing VNN with Reactome pathway hierarchy

We designed VNN structure based on the Reactome pathway hierarchy that comprises root biological processes, child-parent pathway relations, and protein-pathway annotations (downloaded in Aug 2019)⁷⁵. To trim the scale of the neural network and prevent overfitting, we adopted two hyperparameters to filter Reactome pathways: (i) minimal pathway size (values for tuning: 5, 20) and (ii) root biological process (values for tuning: 'gene expression', 'immune system', 'metabolism', 'signal transduction', and all possible combinations among the four, 15

values in total). We selected the four processes due to their broad coverage and direct involvement in cellular mechanism of toxicity. Each pathway is coded as a hidden module with fixed number of neurons. For a pathway p , the number is defined by:

$$N_p = \text{round}\left[1 + (N_{max} - 1) * \frac{\log S_p/S_{min}}{\log S_{max}/S_{min}}\right]$$

where S_p denotes the size of p , S_{min} and S_{max} denote the minimal and maximal size of a pathway in the VNN, respectively, and N_{max} ($= 20$) denotes the maximal number of neurons for a hidden module. As a result, hidden modules of larger pathways are assigned with more neurons to capture potentially more complex responses.

Under the Reactome hierarchy, the VNN model of DTox starts from an input layer containing 361 protein features, which are connected to lowest-level hidden modules by protein-pathway annotations. The connections to a hidden module of pathway p are encoded by a weight matrix \mathbf{W}_p with dimensions $N_p * N_{protein}$, where N_p denotes the hidden module size, and $N_{protein}$ denotes the number of input proteins annotated with p . With \mathbf{W}_p , input vector \mathbf{x}_p is transformed to output vector \mathbf{y}_p via:

$$\mathbf{y}_p = \text{ReLu}[\mathbf{x}_p \mathbf{W}_p^T + \mathbf{b}_p]$$

where \mathbf{b}_p is a bias vector. The hidden modules are then interconnected by child-parent pathway relations until root biological processes are reached. Finally, the root biological processes are connected to an output layer containing the assay outcome. The connections to the output layer are encoded by a weight matrix \mathbf{W}_r with dimensions $1 * N_r$, where N_r denotes the sum of root hidden module sizes. The final output y_r is computed as:

$$y_r = \text{Sigmoid}[\mathbf{x}_r \mathbf{W}_r^T + b_r]$$

where the Logistic Sigmoid function converts layer inputs to an output score between 0 and 1 (i.e., the predicted outcome probability). In addition, we adopted the idea of auxiliary layers from DCell⁹⁶ to prevent gradients from vanishing in the lower hierarchy, and to facilitate the learning of

new patterns from individual pathways. Specifically, output vector of hidden module \mathbf{y}_p is transformed to an auxiliary scalar y'_p via:

$$y'_p = \text{Sigmoid}[\mathbf{y}_p \mathbf{W}'_p + b'_p]$$

where \mathbf{W}'_p denotes the weight matrix with dimensions $1 * N_p$. The auxiliary scalars from all hidden modules are then evaluated in a loss function along with the final output:

$$BCELoss(y_r, y) + \alpha \sum_p \beta_p BCELoss(y'_p, y) + \lambda \|W\|_2$$

The auxiliary factor α is a hyperparameter of the VNN model (values for tuning: 0.1, 0.5, 1), balancing between root and auxiliary loss terms. β_p serves as the adjustment factor for auxiliary loss terms from pathway p , being computed as the inverse number of pathway count within the corresponding hidden layer. Therefore, pathways higher in the hierarchy exhibit greater contribution to the loss function as pathway count decreases dramatically along the hierarchy. λ ($= 1e^{-4}$) is the coefficient for L_2 regularization.

3.2.3 Learning optimal DTox model for Tox21 assay outcome prediction

Each dataset is split into learning and validation sets by ratio of 4:1. During model training, the learning set is further split into training and testing sets by ratio of 7:1. The purpose of the split is to set aside an independent testing set for overfitting assessment during model training. At every epoch, forward and backward propagation are performed on the training set for deriving gradients of model parameters. The parameters are then optimized by Adam algorithm with mini-batch size of 32. At the end of every epoch, loss function is evaluated on the testing set for assessing overfitting and determining whether the early stopping criterion has been met (testing loss has not decreased for P epochs, where P represents the “patience” hyperparameter and is set 20 in this study). Model training stops after 200 epochs or if the early stopping criterion has been met (in our experience, the early stopping criterion is often met long before 200 epochs).

As mentioned above, the VNN model of DTox has three hyperparameters: minimal pathway size, root biological process, and the auxiliary factor α . To find the optimal setting for

each assay, we adopted grid search and implemented all possible hyperparameter combinations to train DTTox models (90 combinations in total, listed in Table 3.1). We evaluated each trained model by computing the loss function on the whole learning set, then identified the optimal model that minimizes learning loss. Finally, the held-out validation set was used to evaluate the performance of the optimal DTTox model and compare with other machine learning models. We adopted two performance metrics for the task: area under the ROC curve (AUROC) and balanced accuracy. We computed the 95% confidence interval (CI) of metrics using bootstrapped samples from predicted outcome probabilities. On average, the bootstrapped samples contain 63.3% of unique original samples. The performance of two methods is significantly different if their CIs do not overlap. Three machine learning models were considered for performance comparison: (i) A fully-connected multi-layer perceptron model with the same number of hidden layers and neurons as optimal DTTox model, (ii) an optimal random forest model derived from tuning of six hyperparameters (“n_estimators”, “criterion”, “max_features”, “min_samples_split”, “min_samples_leaf”, and “bootstrap”) by grid search (2800 combinations in total, listed in Table 3.1), and (iii) an optimal gradient boosting model derived from tuning of five hyperparameters (“n_estimators”, “max_depth”, “learning_rate”, “subsample”, and “min_child_weight”) by grid search (3000 combinations in total, listed in Table 3.1).

In addition, shuffling analysis was performed to assess the influence of pathway knowledge and hierarchy on DTTox performance. Three distinctive layouts were considered for performance comparison: (i) An alternative DTTox model built under shuffled Reactome ontology hierarchy while the shuffle preserves the number of children for each parent pathway and the number of connections between hidden layers (Suppose a parent pathway is connected to three children in the original DTTox, two in layer i and one in layer j . By hierarchy shuffling, the parent will be connected to two pathways sampled from layer i and one pathway sampled from layer j . This shuffling strategy ensures that the resulting DTTox model is still consecutively connected from input to output layer.) (ii) an alternative DTTox model built with shuffled input target profile (the input values are shuffled among features). (iii) an alternative DTTox model built with shuffled assay

outcome as negative control (the outcome labels are shuffled among compounds within the learning set).

DTox visible neural network		
Hyperparameter name	Hyperparameter description	Search values
min_pathway_size	minimal size of pathway to be included in the network	5, 20
root_pathway	root biological process to be included in the network	"gene expression", "immune system", "metabolism", "signal transduction", and all possible combinations among the four processes, 15 values in total
alpha	coefficient for auxiliary loss term	0.1, 0.5, 1
Random forest		
Hyperparameter name	Hyperparameter description	Search values
n_estimators	The number of trees in the forest	50, 100
criterion	The function to measure the quality of a split	"gini", "entropy"
max_features	The proportion of features to consider when looking for the best split	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
min_samples_split	The minimum number of samples required to split an internal node	2, 5, 8, 11, 14, 17, 20
min_samples_leaf	The minimum number of samples required to be at a leaf node	1, 6, 11, 16, 21
bootstrap	Whether bootstrap samples are used when building trees	True, False
Gradient boosting		
Hyperparameter name	Hyperparameter description	Search values
n_estimators	The number of gradient boosted trees	50, 100
max_depth	The maximum tree depth for base learners.	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
learning_rate	The boosting learning rate	0.1, 0.01, 0.001
subsample	The subsample ratio of the training instance.	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
min_child_weight	The minimum sum of instance weight(hessian) needed in a child	1, 6, 11, 16, 21

Table 3.1 Hyperparameter tuning of classification algorithms

3.2.4 Interpreting optimal DTox model by layer-wise relevance propagation

Layer-wise relevance propagation¹⁰³ (LRP) is a model interpretation tool for deep neural networks. Through backward propagation, LRP assigns each neuron a share of the network output and redistributes it to its predecessors in equal amounts until the input layer is reached. The propagation procedure ensures that relevance conservation is an inherent property of LRP. To implement LRP, we adopted two local propagation rules: Generic rule and input-layer rule¹⁰⁴.

Generic rule was applied to relevance propagation of the hidden neurons. For two connected neurons j and k from a child-parent pathway pair, the forward propagation of VNN follows

$$a_k = \text{ReLu}(\sum_j a_j w_{jk} + b_k)$$

where a_k denotes the activation of neuron k . The generic rule propagates relevance between them as:

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma w_{jk}^+)}{\varepsilon \cdot SD[(w_{jk} + \gamma w_{jk}^+)_{jk}] + \sum_j a_j \cdot (w_{jk} + \gamma w_{jk}^+)} R_k$$

where γ and ε are two hyperparameters of the rule. γ (values for tuning: 0.001, 0.01, 0.1) controls the contribution of positive weights in relevance propagation. Increasing the value of γ can marginalize neurons with negative weights and decrease the variance of relevance across neurons, and thus may lead to more stable interpretation results. ε (values for tuning: 0.001, 0.01, 0.1) absorbs relevance from neurons with weak or contradictory weights. Increasing the value of ε can give prominence to a few neurons with high weights, and thus may lead to more sparse interpretation results.

Input-layer rule was only applied to relevance propagation of the input protein features. For a protein feature i and its connected neuron j from a lowest-level pathway, the input-layer rule propagates relevance between them as:

$$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$$

where $l_i (= 0)$ and $h_i (= 1)$ are the lower and upper bound of input feature values.

3.2.5 Identifying significant VNN paths for explaining toxicity outcome of compounds

After relevance of each neuron is assigned via LRP, a relevance score is computed for each pathway by summing the relevance scores of its neurons. An observed score is then computed for each VNN path connecting input protein feature to output assay outcome as:

$$S_{path} = \sum_{p \in path} \log R_p^+$$

where p denotes a protein or pathway along the path. The relevance scores are converted to non-negative values, as we are only interested in the proteins or pathways that are more likely to result in a toxicity outcome. The log transformation is adopted to adjust the scale of relevance

scores from different layers, as the number of pathways decreases dramatically along the hierarchy.

To assess the significance of each observed path score, we employed a permutation-based strategy to derive the null distribution. Specifically, we shuffled the outcome label of each Tox21 dataset, then re-trained random DTox models using the same hyperparameter setting as previously trained optimal model. The procedure was repeated $n = 200$ times, a balance between sample size and running time. Scores derived from the random DTox models comprise the null distribution for each observed path score, and thus the empirical p -value can be computed as

$$S_{path} = \sum_{i=1}^N I(S_{path-i} \geq S_{path})/n$$

We used the false discovery rate (FDR) to perform multiple testing correction on all VNN paths, then identified the significant paths (FDR < 0.05) for each active compound.

As mentioned above, DTox's interpretation framework has two hyperparameters: γ and ε from the generic rule. To study the effect of hyperparameter settings on model interpretation, we implemented all possible (9 in total) hyperparameter combinations to identify significant VNN paths for active compounds. We measured the similarity between each pair of settings by the median Jaccard Index among active compounds regarding their identified significant paths.

3.2.6 Comparing DTox against existing interpretation methods regarding rediscovering mechanisms of transcription activation by nuclear receptor

Three interpretation methods were considered for performance comparison regarding the task. The first method serves as a baseline for DTox interpretation framework, in which we randomly sampled the same number of VNN paths for each compound as identified by DTox, from the pool of all possible paths in the network. The performance metric was computed as the proportion of active compounds that were sampled with the "ground truth" VNN path (linking together root process of gene expression, nuclear receptor transcription pathway, and the specific target receptor). The procedure was repeated 1,000 times to account for the stochastic nature of sampling. The average performance and 95% confidence interval were computed and adopted as baseline for DTox.

The second method is widely used for explaining predictions of classification algorithms, namely LIME⁹³. LIME explains predictions by fitting local linear models to approximate the behavior of original model. For each nuclear receptor of interest, we implemented LIME on the optimal random forest model (derived previously from hyperparameter tuning) to explain the predicted outcome of each compound by target feature relevance (our implementation was based on the tutorials in <https://github.com/marcotcr/lime>). The performance metric was computed as the proportion of active compounds that were explained with high relevance regarding the specific target receptor. We adopted two thresholds for defining “high relevance”: (i) Feature relevance for the target receptor is positive (lax threshold). (ii) Feature relevance for the target receptor is above average (strict threshold).

The third method is commonly used for inferring toxicity profile of new compounds, namely Read-across. Read-across does not rely on classification algorithms. Instead, it assigns existing knowledge on source compounds to the query compounds with similar chemical structure. For each nuclear receptor of interest, we extract compounds with known connections (source compounds) from two resources: DrugBank⁶⁹ and ComptoxAI¹⁰⁵. The performance metric was computed as the proportion of active compounds (query) that exhibit similar structure to at least one source compound. Five thresholds of Tanimoto coefficient were adopted to define structural similarity between source and query compound: 0.8, 0.85, 0.9, 0.95, and 1.

3.2.7 Processing LINCS dataset for validation of DTox interpretation results

The LINCS dataset¹⁰⁰ contains gene-expression profiles derived from genetic and small-molecule perturbation experiments on a number of cell lines, including MCF-7 (which was used in Tox21’s aromatase assay) and HepG2 (used in Tox21’s mitochondria toxicity assay, PXR agonist assay, and HepG2 cell viability assay). We extracted the profiles induced by active compounds of the four assays in their respective cell line. We removed the profiles that did not pass quality control, then separated the remaining ones into three groups based on dose and time of perturbation (1.11 μ M-24h, 10 μ M-6h, 10 μ M-24h). We used the LINCS level 5 data, which consists of moderated differential expression Z-scores, for the validation analysis.

To assess the differential expression of VNN paths identified for each compound, we first identified differentially expressed genes (DEGs) from the corresponding profile by $|Z| > 2$, as suggested by LINCS. Then, we used Fisher's exact test to examine whether the pathways along each VNN path are enriched for DEGs. A test p -value was computed for each pathway. We used FDR to perform multiple testing correction on all pathways along each path. A VNN path is differentially expressed if all the pathways involved are significantly enriched for DEGs (FDR < 0.05). Finally, we calculated the proportion of differentially expressed paths among the paths identified by DTox (observed proportion) and among all possible paths in VNN (expected proportion).

3.2.8 Processing datasets for analyzing DTox results on HepG2- and HEK293-cytotoxic compounds

We obtained six DrugBank lists from <https://go.drugbank.com/releases/latest#external-links>⁶⁹. Each list contains a number of compounds sharing a particular approval status. We obtained 265 EPA chemical lists from <https://comptox.epa.gov/dashboard/chemical-lists>. Each list contains a number of compounds sharing a particular property.

The NSIDES dataset²³ contains drug-adverse event relations that are derived from FDA reports after adjusting for confounding factors. Each drug-adverse event pair is assigned with a proportional reporting ratio (PRR) score along with its 95% CI, which measures the extent to which the adverse event is disproportionately reported among individuals taking the drug. We manually curated a list of 20 clinical phenotype terms associated with drug-induced liver injury (DILI) and a list of 24 clinical phenotype terms associated with drug-induced kidney injury (DIKI). Drugs associated with each phenotype of interest are identified by the lower bound of 95% CI (> 1). Drugs not associated with each phenotype of interest (negative controls) are identified by both the lower (< 1) and the upper (> 1) bound of 95% CI.

To measure the association between each DILI phenotype and HepG2 cytotoxicity, we calculated the odds ratio and its 95% CI based on a 2*2 contingency table. The same procedure was performed to measure the association between each DIKI phenotype and HEK293

cytotoxicity. We also used Fisher's exact test to evaluate the enrichment of nine cell death-related pathways among the drugs associated with DILI phenotypes. The odds ratio and test P-value were computed for each phenotype-pathway pair. We used FDR to perform multiple testing correction on all phenotype-pathway pairs.

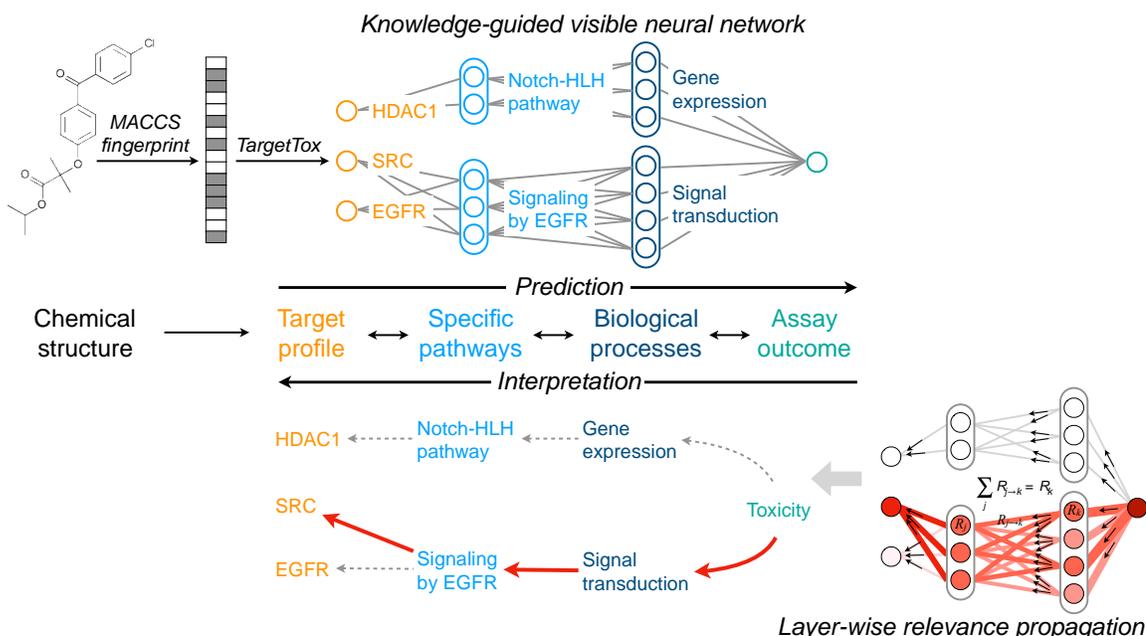


Figure 3.1 Modeling compound response to toxicity assay with DTox

For toxicity prediction, the chemical structure of a compound is quantified using MACCS fingerprint before being converted to target profile by our previously developed method, TargetTox. The target profile is then fed into a VNN, whose structure is guided by Reactome pathway hierarchy. Specific pathways and biological processes are coded as hidden modules with a series of neurons. For model interpretation, the network output is propagated backward onto each neuron as relevance score using the layer-wise relevance propagation technique. A permutation-based strategy is then employed to identify the VNN paths of high relevance. Each path connects a compound to its toxicity outcome via the target protein, specific pathways, and biological process.

3.3 Results

3.3.1 Training DTox for predicting compound response to toxicity assays

The purpose of DTox is to predict the outcome of interest from chemical structure of compounds, and to explain the predicted outcome with activities of proteins and pathways. To train the model, DTox takes in a labeled dataset that specifies the 2D structural representation of each compound (in the form of SMILES string) along with the binary outcome of a screening assay (active or inactive). Since a VNN model typically starts with input layers consisting of gene

or protein features, to fill in the gap, we first quantified the structure of each compound using a 166-bit MACCS fingerprint (each bit represents the answer to a yes/no question regarding chemical structure), then applied our previously developed method, named TargetTox¹⁰², to derive a target profile of each compound. TargetTox was pre-trained on experimentally measured compound-target binding affinities to infer the target binding probability of each compound from its MACCS fingerprint. The derived profile contains 361 target proteins, spanning six functional categories: Enzymes, G protein coupled receptors, catalytic receptors, ion channels, nuclear hormone receptors, and transporters. We designed a VNN structure (Figure 3.1) that connects target proteins (input features) to assay outcomes (output response) via Reactome pathways (hidden modules). By our design, each pathway is represented by 1-20 neurons depending on its size. Connections between input features and the first hidden layer are constrained to follow protein-pathway annotations while the connections among hidden layers are constrained to follow child-parent pathway relations. The incorporation of pathway hierarchy makes DTox models highly interpretable, in contrast to conventional black-box neural network models.

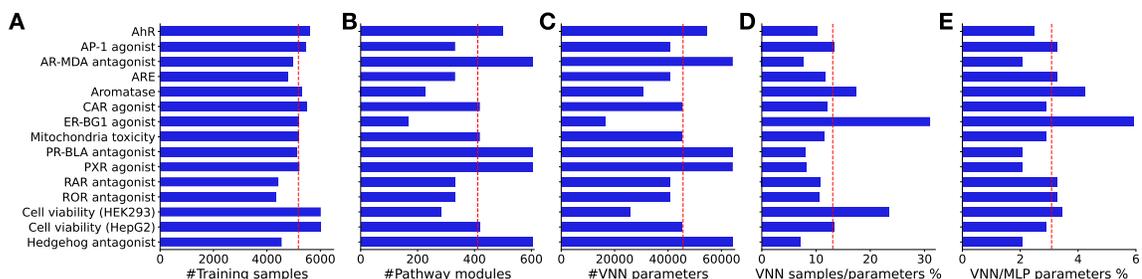


Figure 3.2 Comparison of DTox model statistics

Barplots showing the comparison of different model statistics for 15 toxicity assays: (A) the number of compounds in the training set, (B) the number of hidden pathway modules in the optimal DTox model, (C) the number of trainable parameters in the optimal DTox model, (D) the ratio between number of compounds in the training set versus number of trainable parameters in the optimal DTox model, and (E) the ratio between number of trainable parameters in the optimal DTox model versus the matched MLP model. The dashed red line in each panel represents the average across all 15 assays.

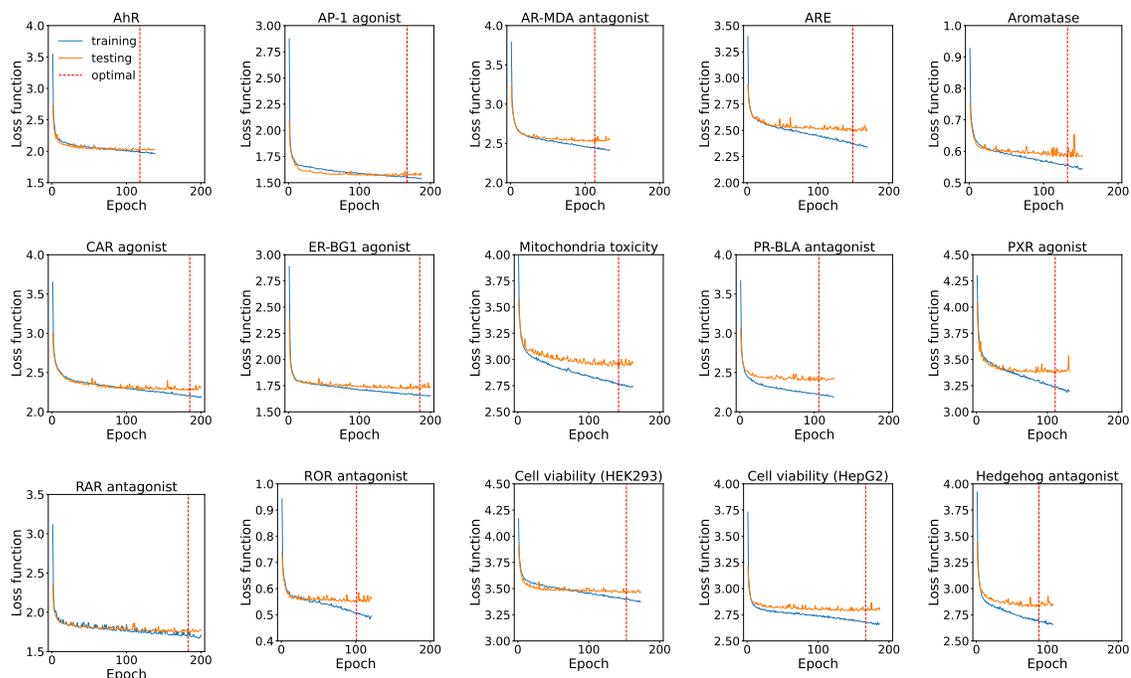


Figure 3.3 Evolution of loss function during learning of optimal DTox model

Line charts showing the evolution of loss function over epochs during learning process of optimal DTox models for 15 toxicity assays. Two types of loss functions are calculated and shown: loss on the training set (blue line, labeled as training) and loss on the testing set (orange line, labeled as testing). The dashed red line in each chart represents the epoch when optimal model is reached. The testing loss does not decrease for 20 consecutive epochs after the optimal point. AhR: aryl hydrocarbon receptor, AP-1: activator protein-1, AR-MDA: androgen receptor in MDA-kb2 AR-luc cell line, ARE: antioxidant response element, CAR: constitutive androstane receptor, ER-BG1: estrogen receptor in BG1 cell line, PR-BLA: progesterone receptor in PR-UAS-bla HEK293T cell line, PXR: pregnane X receptor, RAR: retinoid acid receptor, ROR: retinoid-related orphan receptor.

We trained DTox models on 15 datasets (Table 3.2) from the Tox21 high throughput screening program¹⁷. A DTox model was learned separately for each dataset to predict the active/inactive status of compounds (i.e. screening results of the toxicity assay). On average, each dataset contains 5,178 compounds available for DTox training, including 746 active compounds and 4,432 inactive compounds (Figure 3.2A). To assess model overfitting during the training process, we withheld an independent testing set from each dataset to monitor the evolution of the loss function, and an early stopping criterion to conclude training when overfitting starts to occur. We discovered that while training loss continues to decrease, testing loss stops decreasing after 100-150 epochs for most datasets, a sign of model overfitting (Figure 3.3). Therefore, when overfitting was detected, DTox would conclude training and output the optimal

model when minimal testing loss was reached. On average, the optimal DTox model was learned over 140 ± 16 epochs. We implemented hyperparameter tuning by grid search to derive an optimal model for the prediction of each assay outcome. On average, an optimal DTox model contains 412 hidden pathway modules (Figure 3.2B), and 45,623 neural network parameters (Figure 3.2C). The average ratio between number of training samples versus number of network parameters is 0.13 ± 0.03 (Figure 3.2D), with the estrogen receptor agonist assay model being the highest (0.31) and the hedgehog antagonist assay model being the lowest (0.07). Compared to a conventional multi-layer perceptron (MLP) model, DTox model has far fewer network parameters. On average, the number of network parameters for a DTox only accounts for three percent of the number for a matched MLP (Figure 3.2E).

Tox21 dataset name	Tox21 assay name	Assay target category	Cell line	Total number of compounds	Number of active compounds	Number of inactive compounds
tox21-ahr-p1	AhR	nuclear receptor	HepG2	7008	764	6244
tox21-ap1-agonist-p1	AP-1 agonist	stress response	ME-180	6847	572	6275
tox21-ar-mda-kb2-luc-antagonist-p2	AR-MDA antagonist	nuclear receptor	MDA-MB-453	6199	886	5313
tox21-are-bla-p1	ARE	stress response	HepG2	5973	996	4977
tox21-aromatase-p1	Aromatase	stress response	MCF-7	6627	714	5913
tox21-car-agonist-p1	CAR agonist	nuclear receptor	HepG2	6856	883	5973
tox21-er-luc-bg1-4e2-agonist-p2	ER-BG1 agonist	nuclear receptor	BG1	6481	704	5777
tox21-mitotox-p1	Mitochondria toxicity	stress response	HepG2	6479	1199	5280
tox21-pr-bla-antagonist-p1	PR-BLA antagonist	nuclear receptor	HEK293	6426	829	5597
tox21-pxr-p1	PXR agonist	nuclear receptor	HepG2	6526	1618	4908
tox21-rar-antagonist-p2	RAR antagonist	nuclear receptor	C3H10T1/2	5535	554	4981
tox21-ror-cho-antagonist-p1	ROR antagonist	nuclear receptor	CHO	5409	500	4909
tox21-rt-viability-hek293-p1	Cell viability (HEK293)	cytotoxicity	HEK293	7523	1717	5806
tox21-rt-viability-hepg2-p1	Cell viability (HepG2)	cytotoxicity	HepG2	7526	1133	6393
tox21-shh-3t3-gli3-antagonist-p1	Hedgehog antagonist	developmental toxicity	NIH/3T3	5689	927	4762

Table 3.2 Summary of 15 Tox21 datasets used in the study

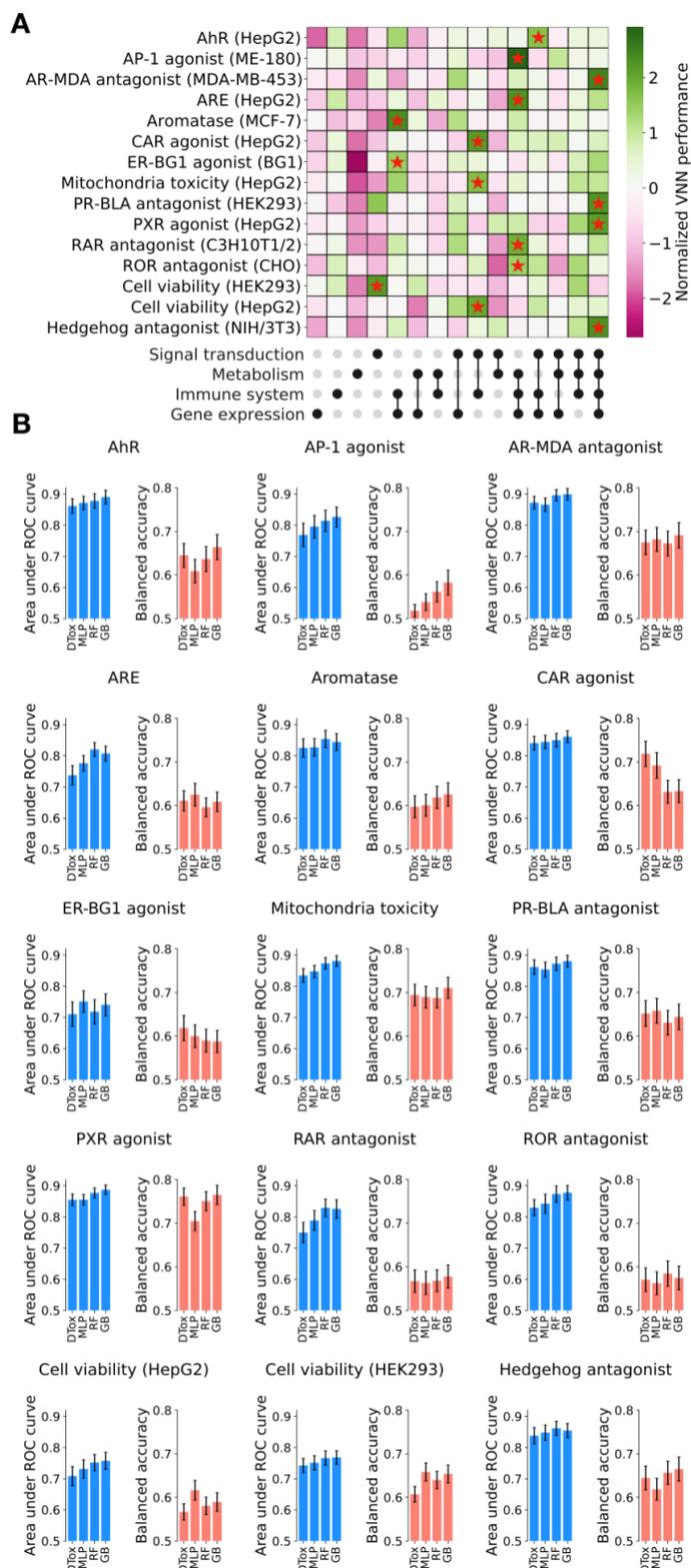


Figure 3.4 Prediction of compound response to 15 toxicity assays

(A) Heatmap showing the training performance of VNN built under different combinations of root biological processes (shown as upset plot at the bottom). To facilitate comparison, the model performance is normalized within each assay using Z-transform. The optimal combination for each assay is highlighted with a red star. The name of each assay is annotated on the left, with name of the assay cell line included in parenthesis. AhR: aryl hydrocarbon receptor, AP-1: activator protein-1, AR-MDA: androgen receptor in MDA-kb2 AR-luc cell line, ARE: antioxidant response element, CAR: constitutive androstane receptor, ER-BG1: estrogen receptor in BG1 cell line, PR-BLA: progesterone receptor in PR-UAS-bla HEK293T cell line, PXR: pregnane X receptor, RAR: retinoid acid receptor, ROR: retinoid-related orphan receptor. (B) Barplot showing the validation performance in all 15 Tox21 datasets. The performance of DTox is compared against three other models: a multi-layer perceptron with the same number of hidden layers and neurons as DTox (MLP), random forest (RF), and gradient boosting (GB). Performance is measured by two metrics: area under ROC curve and balanced accuracy, with error bar shows the 95% confidence interval.

To customize the network structure for prediction of each assay outcome, we made the root biological process a hyperparameter. This means through hyperparameter tuning, we can choose a branch or combination of branches from the Reactome pathway hierarchy that result in the best predictive performance for an assay of interest (Figure 3.4A). For instance, signal transduction pathways alone can deliver the optimal model for HEK293 cell viability assay while additional pathways from the immune system are required for HepG2 cell viability prediction, suggesting a potential role of immune response in HepG2 cytotoxicity. In general, models built with multiple branches perform better than models built with a single branch.

3.3.2 DTox can achieve the same level of performance as complex classification algorithms

We validated the predictive performance of DTox models on held-out validation sets, which on average contain 1,295 compounds per assay. The optimal models of all 15 assays exhibit an area under the ROC curve (AUROC) greater than 0.7 (0.7-0.8: 6 models, 0.8-0.9: 9 models). Similarly, 14 models exhibit a balanced accuracy above 0.55 (0.55-0.65: 9 models, 0.65-0.75: 4 models, > 0.75: 1 model) except for the optimal model of the AP-1 signaling agonist assay. We then compared the optimal performance of DTox against three other classification algorithms (Figure 3.4B). Comparing DTox to a matched MLP model, we observed one assay where DTox significantly outperformed MLP in balanced accuracy (pregnane X receptor agonist), and two assays in the opposite direction (HEK293 and HepG2 cell viability). Comparing DTox to random forest (RF) and gradient boosting (GB), we observed one assay where DTox significantly outperformed both RF and GB (constitutive androstane receptor agonist), and one assay in the

opposite direction (AP-1 signaling agonist). In general, DTox model achieved the same level of predictive performance as these well-established classification algorithms.

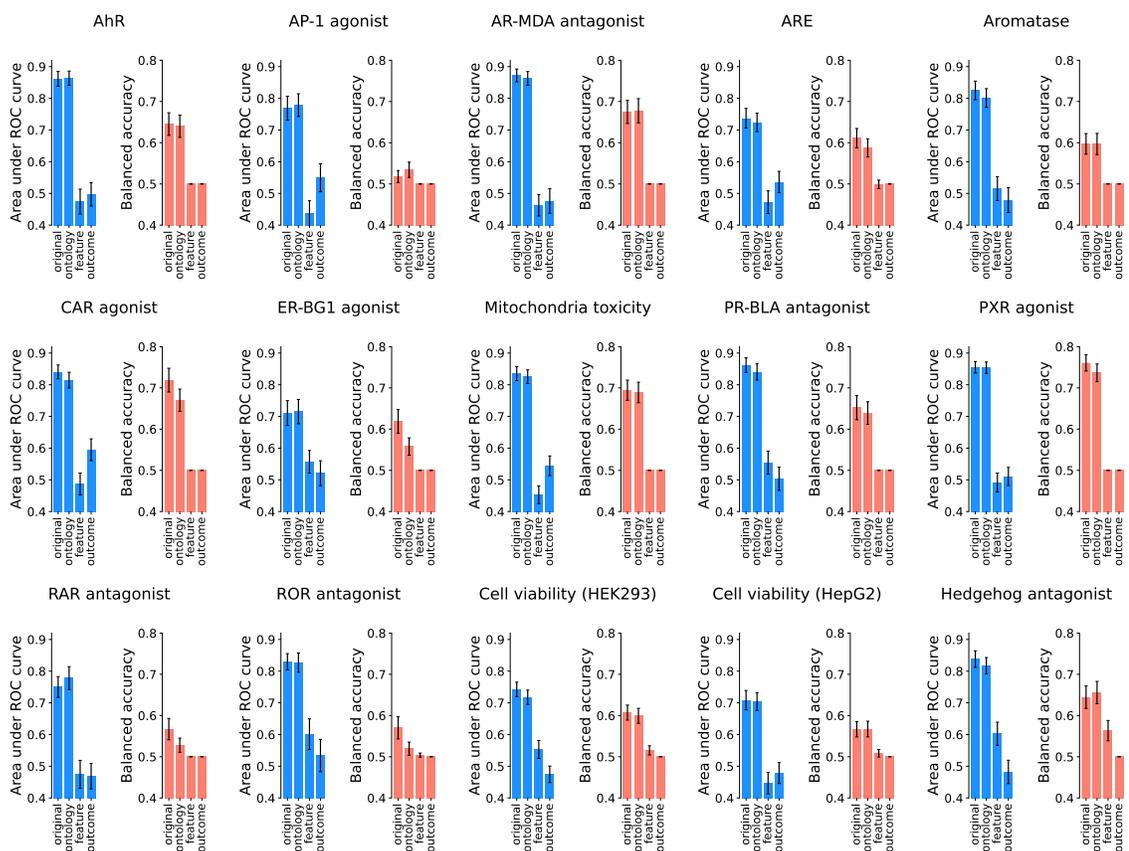


Figure 3.5 Influence of pathway knowledge and hierarchy on predictive performance of DTox

Barplots showing the results of shuffling analysis in all 15 Tox21 datasets. Performance of DTox (original) is compared against three alternative models built with shuffled layouts: (i) an alternative DTox model built under shuffled Reactome ontology hierarchy (ontology), (ii) an alternative DTox model built with shuffled input feature profile (feature). (iii) an alternative DTox model built with shuffled assay outcome as negative control (outcome). Performance on held-out validation set is measured by two metrics: area under ROC curve and balanced accuracy, with error bar shows the 95% confidence interval. AhR: aryl hydrocarbon receptor, AP-1: activator protein-1, AR-MDA: androgen receptor in MDA-kb2 AR-luc cell line, ARE: antioxidant response element, CAR: constitutive androstane receptor, ER-BG1: estrogen receptor in BG1 cell line, PR-BLA: progesterone receptor in PR-UAS-bla HEK293T cell line, PXR: pregnane X receptor, RAR: retinoid acid receptor, ROR: retinoid-related orphan receptor.

To evaluate the degree to which DTox benefits from the incorporation of pathway knowledge, we performed shuffling analysis (Figure 3.5) and compared the predictive performance of original DTox models against alternative models built upon three different layouts: shuffled ontology hierarchy (i.e., child-parent pathway relationships are perturbed), shuffled feature profile (i.e., protein-pathway annotations are perturbed), and shuffled outcome as a

negative control (i.e., input data label is incorrect). We observed that shuffled feature profiles significantly impacted the predictive performance of DTox, as the resulting models exhibited random performance resembling negative controls from the shuffled outcome, suggesting the importance of correct protein-pathway annotations in DTox. By contrast, shuffled ontology hierarchy moderately impacted the predictive performance of DTox, as we observed only two assays where it resulted in significant drop of balanced accuracy (estrogen receptor agonist and retinoid-related orphan receptor gamma antagonist). Notably, the outcomes for both assays are directly related to a specific nuclear receptor transcription pathway, as opposed to other complex outcomes that involve multiple pathways (e.g., mitochondria toxicity, cytotoxicity). That may explain why shuffled ontology hierarchy had a higher impact on these two assays given that connections to the specific pathway would be disturbed in the shuffling.

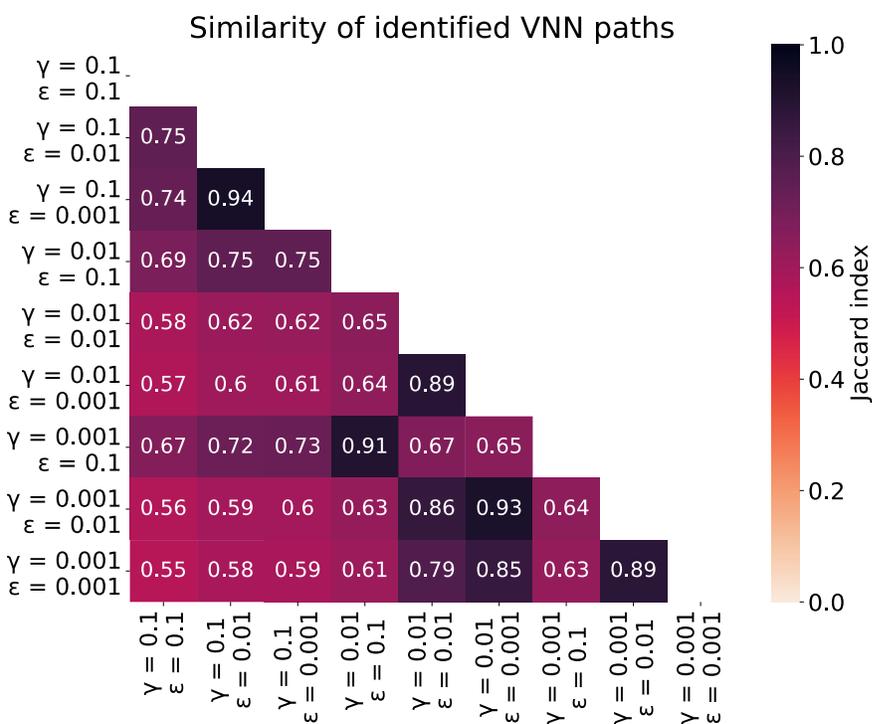


Figure 3.6 Consistency of DTox interpretation across hyperparameter settings

Heatmap showing the similarity of VNN paths identified from nine different hyperparameter settings. The similarity between each pair of setting (annotated in each cell) is measured by the median Jaccard Index among active compounds regarding their identified paths

3.3.3 Development of a DTox interpretation framework for explaining VNN predictions

A fundamental advantage of VNN over other classification algorithms lies in its high interpretability. The incorporation of pathway hierarchy enables us to reason through hidden layers of VNN for mechanistic interpretation. Therefore, we developed a DTox interpretation framework to identify paths from VNN that can explain the toxicity outcome of a compound (Figure 3.1). The framework accepts the derived target profile of each compound along with the trained DTox model that specifies learned weights for each hidden neuron. Each identified path links together a root biological process, its descendant pathway modules, and a target protein feature. The framework has two hyperparameters: γ and ϵ . γ controls the stability of interpretation results while ϵ controls sparsity. We evaluated the effect of hyperparameter settings on identified VNN paths (Figure 3.6). We observed that the set of identified paths exhibits consistently high similarity across distinct hyperparameter settings, as the average Jaccard Index reaches 0.70. Due to the high similarity, we only used the VNN paths identified from one setting ($\gamma = 0.001$, $\epsilon = 0.1$) for the following validation analyses.

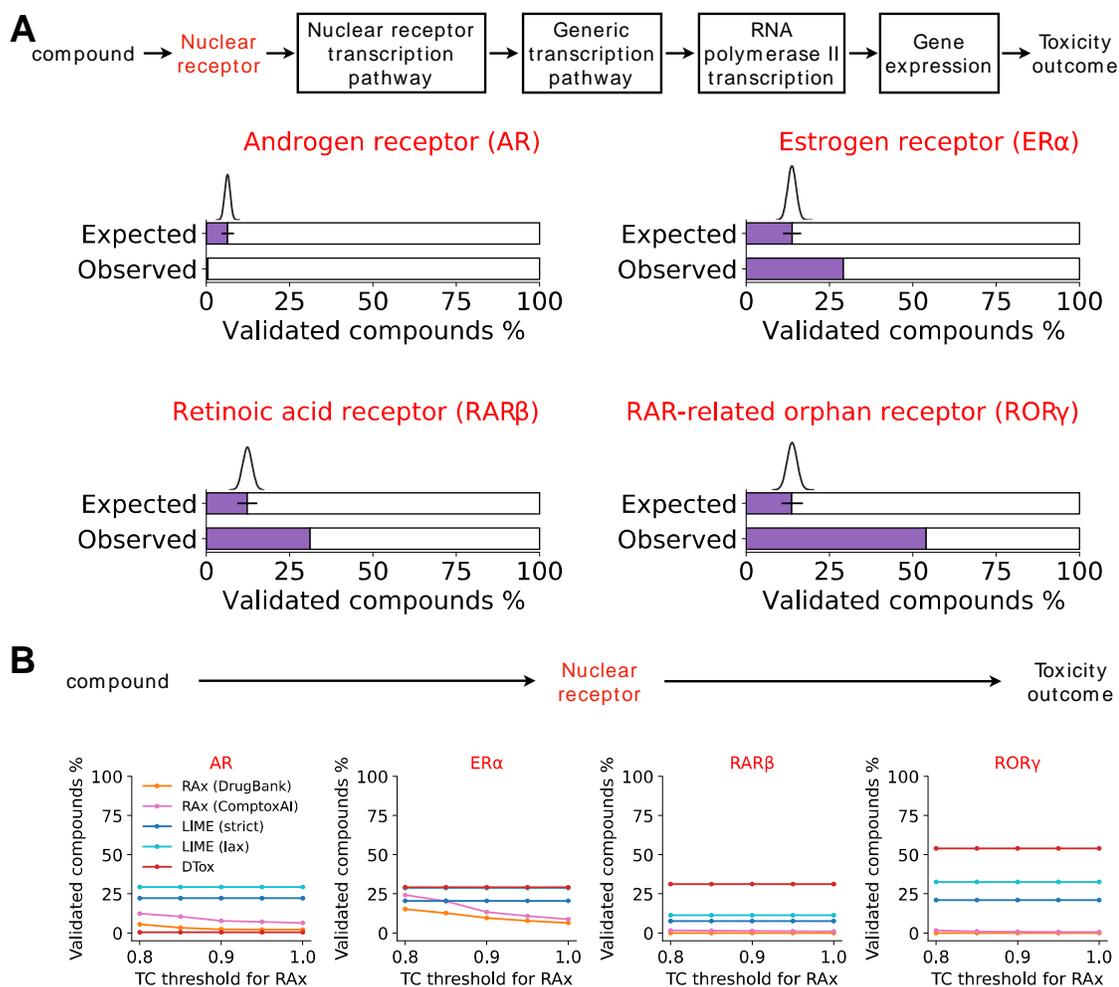


Figure 3.7 Validation of identified VNN paths by known mechanisms

(A) Barplots comparing the observed versus expected proportion of validated compounds in four nuclear receptor assays. The “ground truth” VNN path placed at the top represents the known mechanism of transcription activation by nuclear receptor. A compound is considered to be validated by the known mechanism if the ground truth path is identified by DTox. The expected proportion is computed by random sampling, with the histogram and fitted density curve showing the sampled distribution (95% confidence interval shown as error bar). (B) Line charts comparing the proportion of validated compounds (y-axis) among models. Performance of DTox interpretation framework (DTox) is compared against two other methods: Read-across (Rax) with knowledge source from ComptoxAI or DrugBank, LIME with strict or lax threshold for target feature relevance. Rax models were implemented under five different thresholds of Tanimoto Coefficient (TC; x-axis). A compound is considered to be validated if it can be connected to the nuclear receptor of interest.

3.3.4 DTox can rediscover mechanisms of transcription activation by four nuclear receptors

To evaluate whether DTox can rediscover known mechanisms for a toxicity outcome, we looked for “ground truth” from the VNN paths identified for four nuclear receptor assays:

Androgen receptor antagonist, estrogen receptor agonist, retinoic acid receptor antagonist, and

retinoid-related orphan receptor gamma antagonist. Each of the four assays measures compound response to a specific nuclear receptor transcription pathway. Therefore, we established ground truth as the VNN path that links together the root process of gene expression, nuclear receptor transcription pathway, and the specific target receptor (AR, ER α , RAR β , ROR γ ; Figure 3.7A). In three of the four nuclear receptor assays, our framework was able to identify the ground truth path for at least 29% of all active compounds (ER α : 29%, RAR β : 31%, ROR γ : 54%). Comparing to the expected baseline performance from identifying by chance, our framework improved the proportion by at least twofold.

We also compared the interpretation performance by DTox against two state-of-the-art methods; namely Local Interpretable Model-Agnostic Explanations (LIME), a popular interpretation method for explaining predictions of classification algorithms, and Read-across (RAx), a similarity-based inference technique commonly used in the field of toxicology. Note that neither method provides a mechanism for incorporating pathway knowledge. As a result, they can only connect active compounds to toxicity outcome via target proteins while DTox can provide sample-level explanations linking compounds, target proteins, pathways, and toxicity outcomes. Nevertheless, in three of the four nuclear receptor assays (ER α , RAR β , and ROR γ), DTox exhibits the best interpretation performance while the other two methods display major methodological shortcomings (Figure 3.7B). Specifically, interpretation performance by LIME is dependent on the adopted threshold for feature relevance, as a stricter threshold can significantly deteriorate the performance (e.g., ER and ROR γ). Inference by Read-across, on the other hand, is heavily dependent on the knowledge source, as well as the adopted threshold for similarity measurement. When little existing knowledge could be extracted for the target of interest, Read-across would suffer from poor performance (e.g. RAR β and ROR γ). Despite the strong performance in general, DTox failed to identify the ground truth path for AR antagonists (Figure 3.7A&B). Instead, DTox interpretation linked AR antagonists to target proteins such as integrins, protein kinase A, or protein tyrosine phosphates, which have been shown to regulate the function

of AR^{106, 107}. One possible explanation is that AR antagonists could interact with a variety of off-targets, making it difficult for DTox to aggregate the signal on a single receptor.

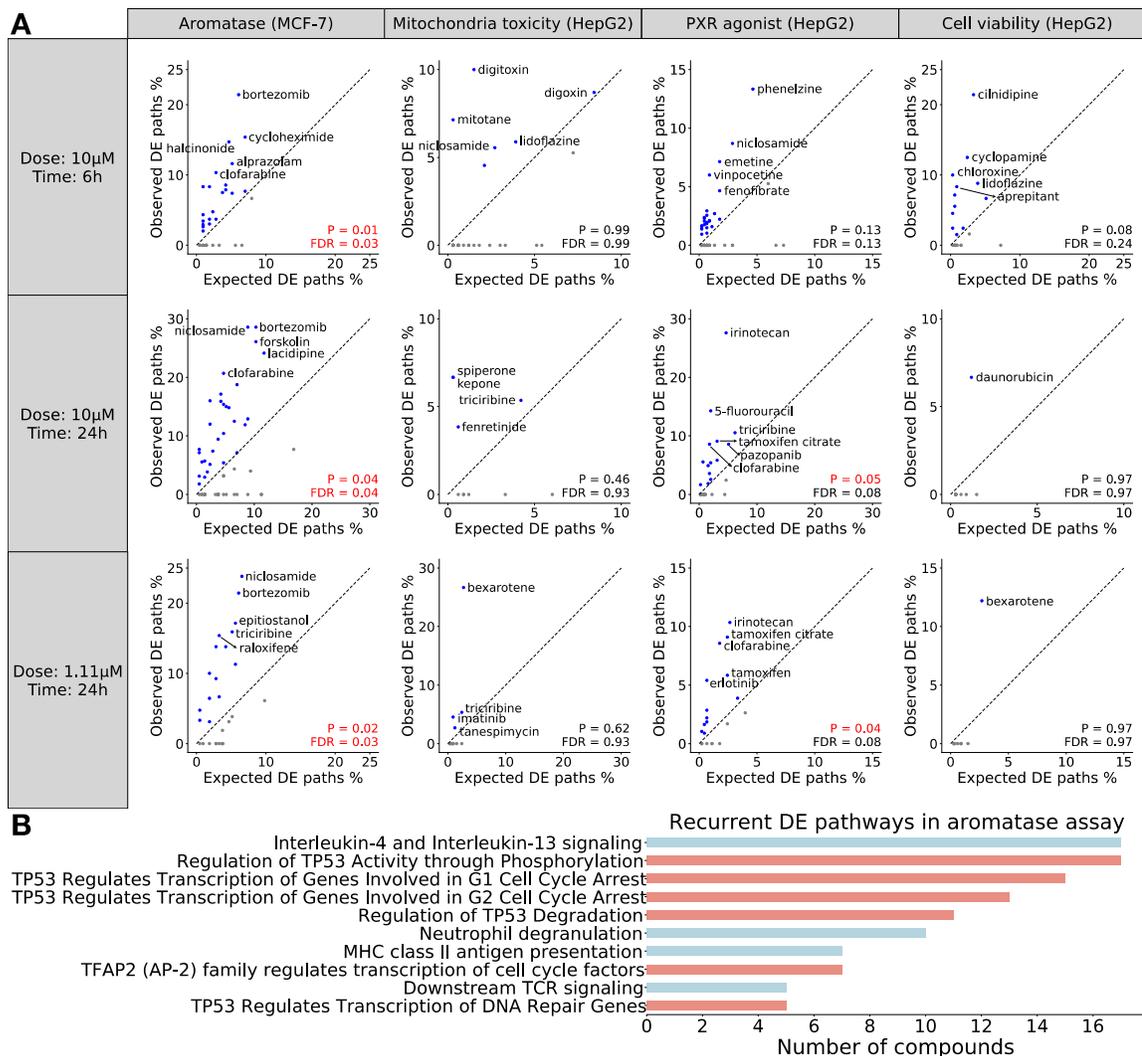


Figure 3.8 Validation of identified VNN paths by differential expression

(A) A VNN path is considered to be differentially expressed (DE) if all pathways along the path is enriched for DE genes from the matched LINCX experiment. The validation analysis is performed for four assays (columns) in three different dose-time groups (rows), with each scatter plot comparing the observed versus expected proportion of DE paths for a single group. The observed proportion is computed with VNN paths identified for each compound while the expected proportion is computed with all possible VNN paths. A Wilcoxon signed-rank test is employed to examine whether the average observed proportion of each group is significantly higher than the average expected proportion (P-value and FDR shown at the bottom right). The diagonal is shown as black dashed line, with compounds in the upper triangle (observed > expected) shown in blue and compounds in the lower triangle (observed < expected) shown in grey. Compounds with the top five observed proportion in each group are annotated with their names. (B) Barplot showing the DE VNN paths that are recurrently identified for at least five aromatase inhibitors. Each VNN path is named after its lowest-level pathway. Paths that contain the “transcriptional regulation by TP53” pathway is highlighted after its lowest-level pathway. Paths that contain the “transcriptional regulation by TP53” pathway is highlighted in salmon while the remaining paths are colored in cyan.

3.3.5 *DTox can recapitulate cellular activities induced by aromatase inhibitors and pregnane X receptor agonists*

To evaluate whether DTox can recapitulate cellular activities induced by active compounds, we studied the differential expression of VNN paths identified for four assays: Aromatase inhibitor, mitochondria toxicity, pregnane X receptor agonist, and HepG2 cell viability. In total, we obtained the gene expression profile measured from 321 LINCS experiments in which an active compound was used to treat the assay cell line (121 experiments for aromatase inhibitor assay, 54 for mitochondria toxicity assay, 101 for pregnane X receptor agonist assay, and 45 for HepG2 cell viability assay). Of all 321 experiments, we found 161 (50%) cases where DTox's interpretation framework was able to identify at least one differentially expressed VNN path. On average, $3.8 \pm 0.6\%$ of VNN paths identified by our framework were found to be differentially expressed, significantly higher than the expected proportion by chance ($2.4 \pm 0.3\%$, $P = 2.5e^{-3}$). We then performed the comparison separately by assay and dose-time combinations (Figure 3.8A). In the aromatase inhibitor assay, our framework outperformed the expected proportion across all three dose-time combinations. In the pregnane X receptor agonist assay, our framework outperformed the expected proportion among the two groups of experiments conducted 24 hours after treatment. In the HepG2 cell viability assay, although no overall difference was detected among experiments conducted 6 hours after treatment, our framework was still able to identify a relatively high proportion of differentially expressed VNN paths for individual compounds such as cilnidipine (21.4%), cyclopamine (12.5%), and chloroxine (10%).

Based on the results of differential expression analysis, induced cellular activities appear to be more consistent among aromatase inhibitors compared to the other three assays, as we discovered ten differentially expressed VNN paths that are recurrently identified for at least five aromatase inhibitors (Figure 3.8B). By contrast, we only discovered one such VNN path for the other three assays combined. Interestingly, "transcriptional regulation by TP53" and its descendant pathways are involved in six of the ten discovered VNN paths, suggesting a potential mechanism for regulation of aromatase by p53 in the MCF-7 aro estrogen responsive element

(MCF-7aro/ERE) breast cancer cell line, a finding supported by a previous study¹⁰⁸. In addition to p53, interleukin-4 and interleukin-13 also appear to play an important role in regulation of aromatase, as the relevant VNN path is linked to 17 aromatase inhibitors by differential expression. This finding is worth further experimental investigation.

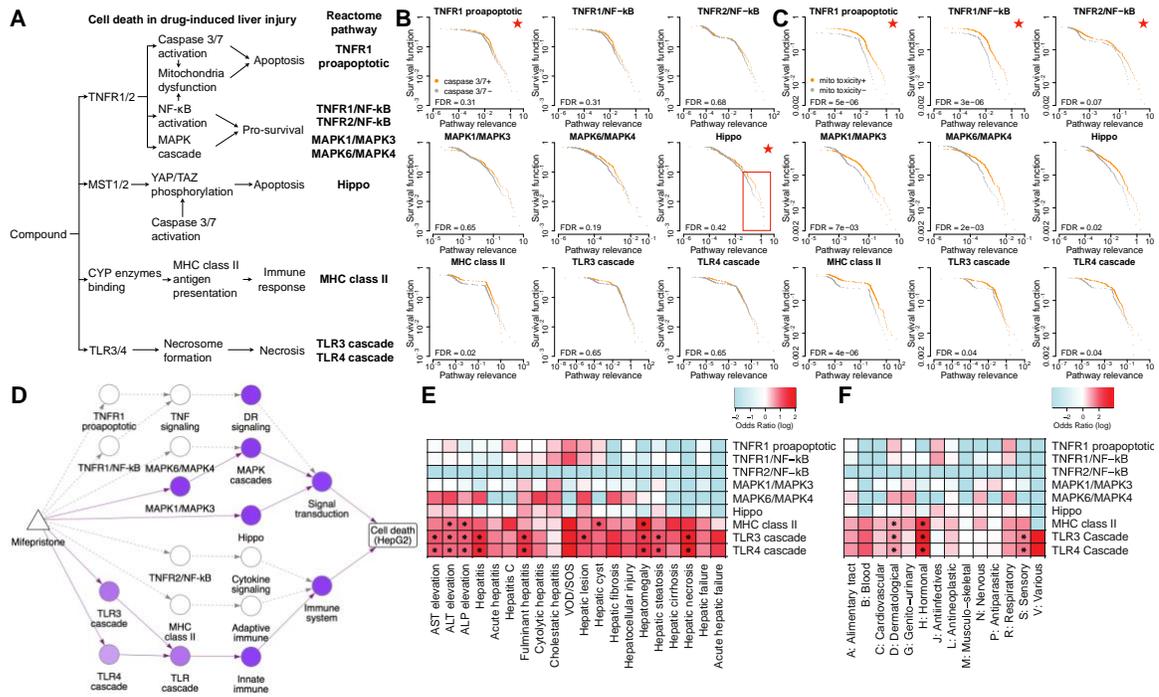


Figure 3.9 In-depth analysis of HepG2 cytotoxicity using identified VNN paths

(A) Established mechanisms for cell death in drug-induced liver injury. Reactome pathways relevant to the mechanisms are identified and used as reference for the analysis. (B and C) Survival plots comparing the pathway relevance scores among active (orange curve) versus inactive (grey curve) compounds of two mechanisms of action assays: caspase 3/7 induction (B) and disruption of the mitochondrial membrane potential (C). Comparisons are made for nine cell death-related pathways, with each plot showing the comparison for a single pathway. Red star at the top right denotes that the pathway is related to the respective mechanism of action. Log-rank test is employed to examine whether the two distributions in each plot are significantly different (FDR value shown at the bottom left). (D) Network diagram showing the simplified DTox structure connecting mifepristone (triangle node) to the HepG2 cytotoxicity (rectangle node) via pathway modules (round nodes). Pathways with relevance score > 0 are colored in purple, with the scale proportional to relevance scale. The VNN paths identified for mifepristone by DTox are shown in solid lines while the rest are shown in dashed lines. (E and F) heatmaps showing the enrichment of nine cell death-related pathways among compounds associated with 20 drug-induced liver injury phenotypes (E) and among compounds of 14 ATC classes (F). Cells are colored based on odds ratio. Fisher's exact test is employed to examine the significance of enrichment (asterisk denotes FDR < 0.05). VOD/SOS: veno-occlusive disease and sinusoidal obstruction syndrome.

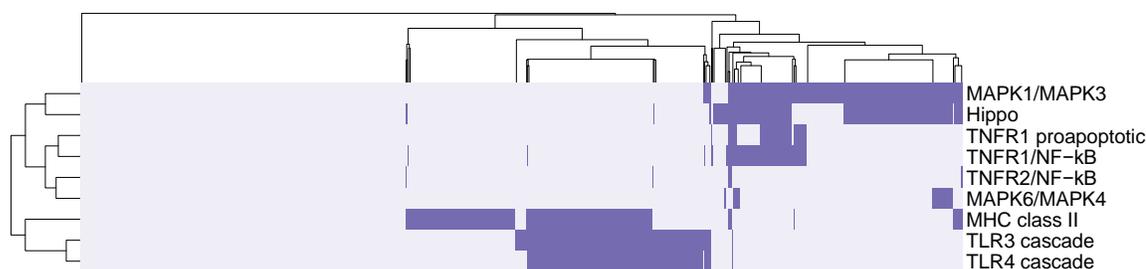


Figure 3.10 Clustering of HepG2-cytotoxic compounds based on cell death-related pathways

Heatmap showing the mapping between 1,120 HepG2-cytotoxic compounds (columns) and nine cell death-related pathways (rows). Hierarchical clustering is performed for both compounds and pathways. Two clusters appear to form as a result. Compounds in the first cluster (top right) are linked to cytotoxicity via apoptosis-related pathways. Compounds in the second cluster (bottom middle) are linked to cytotoxicity via immune-related and necrosis-related pathway.

3.3.6 DTox can differentiate distinctive mechanisms leading to HepG2 cytotoxicity

Next, we sought to explain the compound-induced cytotoxicity in HepG2 cells using VNN paths identified for the HepG2 cell viability assay. A recent review paper¹⁰⁹ summarized four major mechanisms leading to cell death in drug-induced liver injury (DILI): (i) TNFR1/2 mediated apoptosis via caspase activation and pro-survival inhibition, (ii) MST1/2 mediated apoptosis via Hippo signaling, (iii) immune response activation via MHC class II antigen presentation, and (iv) TLR3/4 mediated necrosis (Figure 3.9A). Since the HepG2 cell line was derived from liver tissue, we can use the four mechanisms as a reference for compound-induced cytotoxicity in HepG2 cells. We identified nine Reactome pathways that participate in the four mechanisms (Figure 3.8A). We then mapped HepG2-cytotoxic compounds to the nine Reactome pathways via VNN paths identified by our framework. Of all 1,120 cytotoxic compounds, 707 (63%) compounds are mapped to at least one of the nine cell death-related pathways (Figure 3.10) while the remaining 413 (37%) compounds are mainly linked to HepG2 cytotoxicity via the GPCR, mTOR, and Rho GTPase signaling pathways (Figure 3.11). We performed hierarchical clustering on the mapping and identified two compound clusters (Figure 3.10). Compounds in first cluster are linked to cytotoxicity via apoptosis while compounds in the second cluster are linked to cytotoxicity via immune activation and necrosis. Nevertheless, we discovered a few compounds that exhibit characteristics of both clusters. For instance, according to our framework, mifepristone, a medical abortion drug, causes cytotoxicity in HepG2 cells by activating both apoptosis (via

MAPK1/MAPK3 signaling and Hippo signaling) and necrosis (via TLR3 and TLR4 cascade), a finding supported by previous studies¹¹⁰⁻¹¹² (Figure 3.9D). In addition, our framework was able to link mifepristone with its therapeutic target—the glucocorticoid receptor—via PTK6 signaling). The other therapeutic target of mifepristone—the progesterone receptor—is not in the VNN.

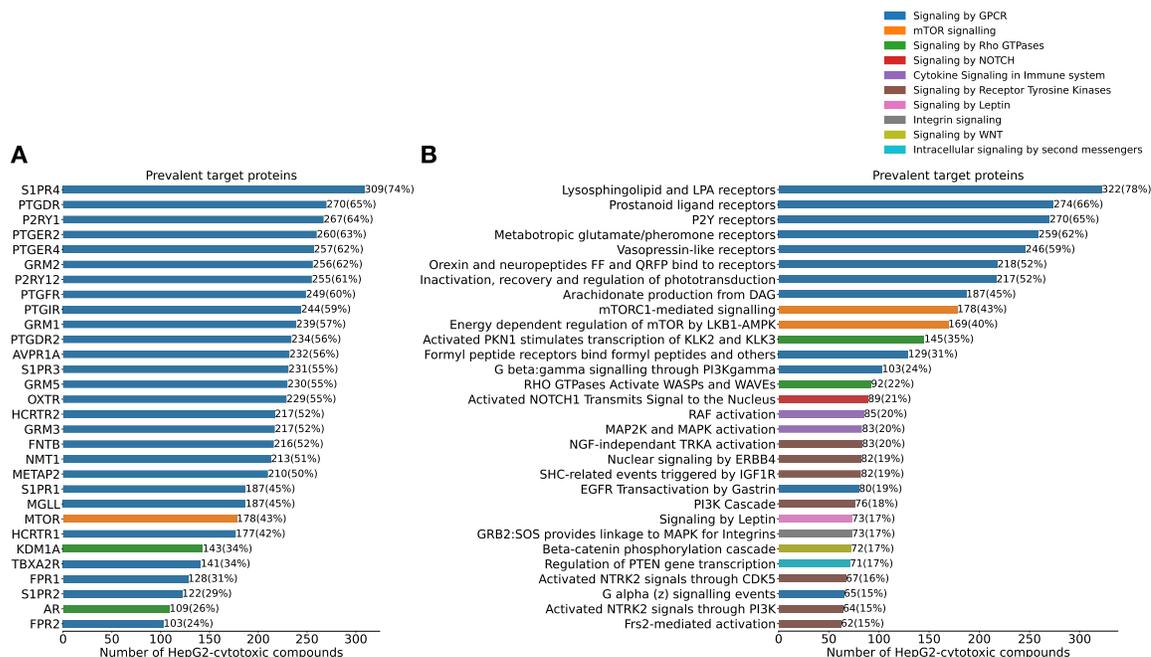


Figure 3.11 Summary of VNN paths identified for 413 cytotoxic compounds not mapped to cell death-related pathways

Barplots showing the top 30 most prevalent target proteins (A) and lowest-level pathways (B) identified for the 413 cytotoxic compounds that cannot be mapped to the nine cell death-related pathways. Each bar is colored by the general category of the target protein or lowest-level pathways it presents, with the color legend shown at top right.

To evaluate whether DTox can differentiate distinctive mechanisms leading to HepG2 cytotoxicity, we looked for concordance between the assigned pathway relevance and screening results from two mechanism of action assays (included in the Tox21 datasets). The first assay we studied measures caspase 3/7 induction in HepG2 cells. Caspase 3 and caspase 7 are key executioners of apoptosis¹¹³. They are involved in TNFR1/2 mediated apoptosis, and YAP/TAZ phosphorylation of Hippo signaling¹¹⁴ (Figure 3.9A). Accordingly, we compared the assigned relevance scores between caspase 3/7+ and caspase 3/7- compounds regarding TNFR1-induced proapoptotic signaling and Hippo signaling (Figure 3.9B). Overall, we did not observe significantly higher relevance among caspase 3/7+ compounds regarding the two signaling pathways (FDR =

0.31 and 0.42, respectively). However, for Hippo signaling, we did observe higher pathway relevance among caspase 3/7+ compounds above the 90th percentile of two distributions (highlighted in Figure 3.9B), hence a partial agreement between assigned pathway relevance and caspase 3/7 induction screening. By contrast, the pattern among top-ranked compounds was not observed in other cell death-related pathways except for MHC class II antigen presentation (FDR = 0.02; Figure 3.9B), suggesting a potential role of caspase 3/7 in MHC class II antigen presentation, a finding worth of further investigation.

The second assay we studied measures disruption of the mitochondrial membrane potential (MMP). MMP is a key indicator of mitochondrial activity as it is required for ATP synthesis. Disruption of MMP can lead to release of cytochrome c, which in turn amplifies the apoptosis signal¹¹⁵. The downstream effectors of TNFR1/2, including caspase activation and inhibition of NF- κ B activation, can cause disruption of MMP^{115, 116} (Figure 3.9A). Accordingly, we compared the assigned relevance scores between MMP-disrupting and nondisruptive compounds regarding TNFR1-induced proapoptotic signaling, TNFR1-induced NF- κ B signaling, and TNFR2-induced NF- κ B signaling (Figure 3.9C). We observed significantly higher relevance among MMP-disrupting compounds regarding TNFR1-induced proapoptotic and TNFR1-induced NF- κ B signaling (FDR = $5e^{-6}$ and $3e^{-6}$, respectively). And the pattern of higher relevance is consistent across all percentiles of two distributions, hence an agreement between assigned pathway relevance and MMP disruption screening. By contrast, the pattern was not observed in other cell death-related pathways except for MHC class II antigen presentation (FDR = $4e^{-6}$; Figure 3.9C), suggesting the potential involvement of mitochondria in antigen presentation, a finding supported by previous work¹¹⁷.

3.3.7 Interpretation of HepG2 cytotoxicity links clinical phenotypes of DILI to TLR3/4 mediated necrosis

We also sought to explain 20 clinical phenotypes of DILI using the derived mapping between compounds and nine cell death-related pathways. For each DILI phenotype, we identified the enriched pathways among its associated compounds (Figure 3.9E). We observed a

disproportionate prevalence of high odds ratio in the two necrosis-related pathways (TLR3 and TLR4 cascade signaling) across almost all DILI phenotypes, with hepatic necrosis, hepatitis, and hepatic fibrosis being the three highest. In total, nine phenotypes are significantly enriched for TLR3/4 mediated necrosis (FDR < 0.05). By contrast, only four phenotypes are significantly enriched for immune activation via MHC class II antigen presentation while no phenotype is significantly enriched for Hippo signaling or TNFR1/2 mediated apoptosis. These results suggest that TLR3/4 mediated necrosis is a common cause for clinical phenotypes of DILI, a finding supported by previous studies^{118, 119}.

Similarly, we identified the enriched pathways among compounds of 14 Anatomical Therapeutic Chemical (ATC) classes (Figure 3.9F). Each ATC class represents a group of drugs that act on a specific organ or system. We found three classes (hormonal, sensory, and dermatological) significantly enriched for TLR3/4 mediated necrosis, and two classes (hormonal and dermatological) significantly enriched for immune activation via MHC class II antigen presentation.

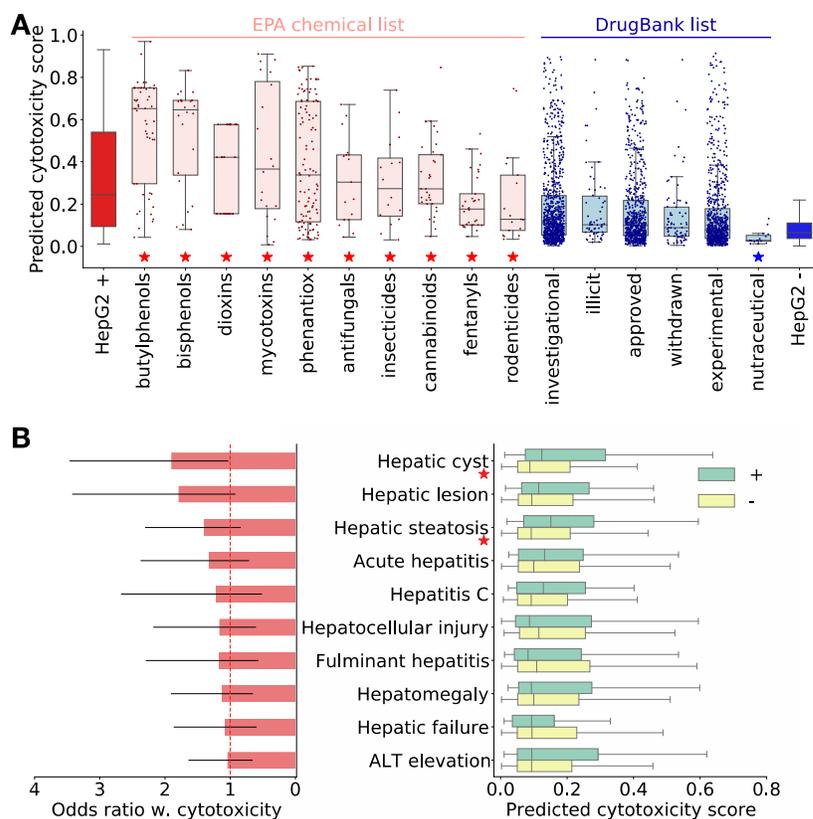


Figure 3.12 Application of predicted HepG2 cytotoxicity score among DSSTox compounds
(A) Boxplot showing the distribution of predicted HepG2 cytotoxicity scores among positive controls (leftmost box in red), ten EPA chemical lists (boxes in light red), six DrugBank lists (boxes in light blue), and negative controls (rightmost box in blue). Mann-Whitney U test is employed to examine whether the cytotoxicity scores of each list exhibit no significant difference from the positive controls (red star above list name), or no significant difference from the negative controls (blue star above list name). **(B)** Boxplot on the right compares the predicted HepG2 cytotoxicity scores among drugs associated with clinical hepatic phenotypes (green box) versus negative controls (yellow box), while barplot on the left shows the odds ratio between HepG2 cytotoxicity and each phenotype (95% confidence interval shown as error bar). Results for ten phenotypes with odds ratio > 1 are shown in the plot. Mann-Whitney U test is employed to examine whether the drugs associated with each phenotype are predicted with higher cytotoxicity scores than the negative controls (red star next to the phenotype name denotes $P < 0.05$).

3.3.8 DTox can be applied to a wide range of chemicals other than drugs

Finally, to demonstrate the applicability of DTox among a broader spectrum of chemicals, we implemented the optimal DTox models of two cell viability assays (HepG2 and HEK293) to predict the probability of cytotoxicity for 708,409 compounds from distributed structure-searchable toxicity (DSSTox)¹²⁰. These compounds provide considerable coverage of the chemical landscape of interest to toxicological and environmental researchers, and have not been screened by the Tox21 project. We first analyzed the predicted HepG2 cytotoxicity by compiled

compound lists from DrugBank regarding drug approval status (Figure 3.12A). We discovered that regardless of approval status, compounds in all six lists exhibited significantly lower predicted HepG2 cytotoxicity than positive controls (active in Tox21 screening). However, only compounds in the nutraceutical list exhibited no significant difference from negative controls (inactive in Tox21 screening). We discovered the same result when analyzing the predicted HEK293 cytotoxicity (Figure 3.13A). These nutraceutical compounds are mostly dietary supplements and food additives that can be taken daily, thus are expected to appear less toxic to human body.

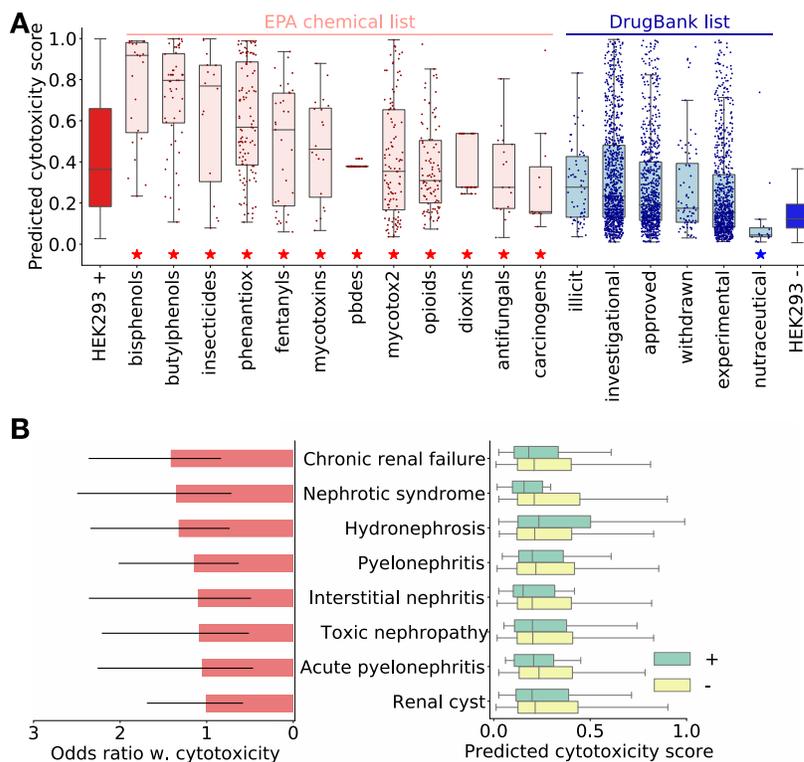


Figure 3.13 Application of predicted HEK293 cytotoxicity score among DSSTox compounds
(A) Boxplot showing the distribution of predicted HEK293 cytotoxicity scores among positive controls (leftmost box in red), 12 EPA chemical lists (boxes in light red), six DrugBank lists (boxes in light blue), and negative controls (rightmost box in blue). Mann-Whitney U test is employed to examine whether the cytotoxicity scores of each list exhibit no significant difference from the positive controls (red star above list name), or no significant difference from the negative controls (blue star above list name). **(B)** Boxplot on the right compares the predicted HEK293 cytotoxicity scores among drugs associated with clinical renal phenotypes (green box) versus negative controls (yellow box), while barplot on the left shows the odds ratio between HEK293 cytotoxicity and each phenotype (95% confidence interval shown as error bar). Results for eight phenotypes with odds ratio > 1 are shown in the plot. Mann-Whitney U test is employed to examine whether the drugs associated with each phenotype are predicted with higher cytotoxicity scores than the negative controls. No significant signal was detected for any phenotypes.

We then analyzed the predicted HepG2 cytotoxicity by compound lists from EPA regarding their chemical properties. Among the 265 chemical lists compiled by EPA, we discovered 12 lists in which compounds exhibited significantly higher predicted HepG2 cytotoxicity than negative controls (inactive in Tox21 screening), and no significant difference from positive controls (active in Tox21 screening). In 10 of the 12 lists (shown in Figure 3.12A), compounds share a common function. Compounds in the other two lists (“casmi2017” and “tscawp”) were compiled together due to joint appearance in contest datasets. Similarly, we discovered 12 such functional lists (shown in Figure 3.13A) when analyzing the predicted HEK293 cytotoxicity. These compounds are either industrial manufacturing products (e.g. bisphenols, dioxins), or lethal to a certain species (e.g. insecticides, rodenticides), thus are expected to appear more toxic to human body. These results also demonstrate that DTox can be applied to a wide range of chemicals other than drugs, including food ingredients, environmental chemicals, industrial chemicals, etc.

3.3.9 HepG2 cytotoxicity scores predicted by DTox can differentiate hepatic cyst compounds from negative controls

To demonstrate the clinical application of DTox, we sought to differentiate DSSTox compounds associated with DILI phenotypes from negative controls using the predicted HepG2 cytotoxicity score (Figure 3.12B). We were able to detect significantly higher predicted scores among the compounds associated with hepatic cyst ($P = 0.015$), as hepatic cyst is the only DILI phenotype showing a significant association with HepG2 cytotoxicity ($OR = 1.90$, 95% CI: 1.04-3.45). Among the remaining 19 DILI phenotypes showing weak or no association with HepG2 cytotoxicity (9 phenotypes with $OR > 1$, 10 with $OR < 1$), we were only able to detect a significant difference for one phenotype: hepatic steatosis ($P = 0.008$). Similarly, we sought to differentiate DSSTox compounds associated with drug-induced kidney injury (DIKI) phenotypes from negative controls using predicted HEK293 cytotoxicity score (Figure 3.13B). Unfortunately, we were not able to detect significant difference for any of the 24 DIKI phenotypes, as none of them exhibits a significant association with HEK293 cytotoxicity (lower bound of 95% OR CI < 1).

	DTox	MLP	Random Forest	Gradient boosting
Average CPU time (in hour)	19.6	1.8	1.1	3.2
Average run time (in hour)	1.72	0.46	0.14	0.15

Table 3.3 Comparison of model efficiency among classification algorithms

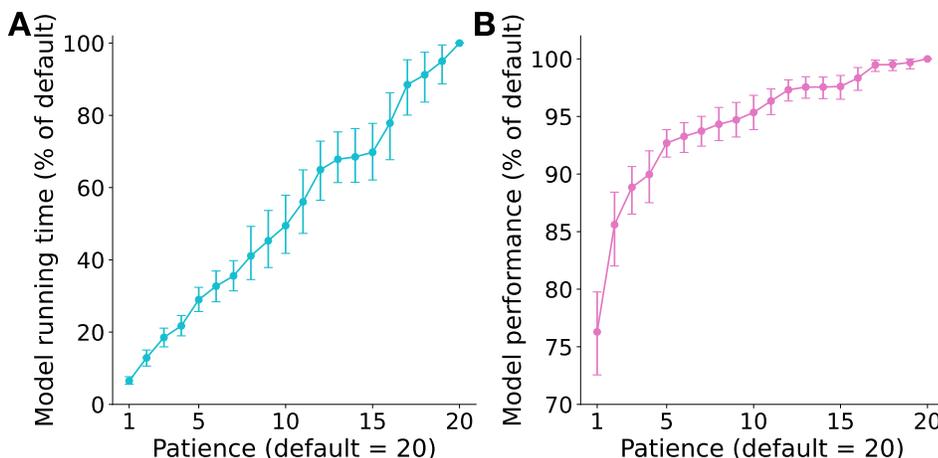


Figure 3.14 Influence of early stopping criterion on DTox model efficiency and performance

(A) Line chart showing the correlation between early stopping criterion (x-axis) and model efficiency (y-axis). The early stopping criterion is quantified by the patience hyperparameter, representing the number of epochs for which testing loss has not decreased before concluding DTox training. The efficiency is measured by relative running time, computed as the ratio between number of running epochs under alternative setting (patience = 1, 2, ..., 20) versus number of running epochs under default setting (patience = 20). Each point represents the average across all 15 Tox21 datasets, with the error bar representing 95% confidence interval. (B) Line chart showing the correlation between early stopping criterion (x-axis) and model performance (y-axis). The performance is measured by relative improvement in testing loss (difference between optimal testing loss and testing loss after epoch one), computed as the ratio between improvement in testing loss under alternative setting (patience = 1, 2, ..., 20) versus improvement in testing loss under default setting (patience = 20).

3.3.10 DTox offers flexibility in balancing between model efficiency and performance

Lastly, we compared the efficiency of DTox model against other classification algorithms. We reported the central processing unit (CPU) time and run time of all algorithms in Table 3.3. On average, it took 1.72 hours for a DTox model to complete training on a single Tox21 dataset (~5000 samples with mini-batch size of 32), three times the duration for MLP, and 12 times the duration for random forest and gradient boosting. As mentioned above, DTox employs an early stopping criterion to conclude training when an optimal model can be detected. In this study, we adopted a conservative stopping criterion for maximizing the predictive performance of derived DTox models. We further tuned the stopping hyperparameter to investigate the flexibility in model

efficiency and performance (Figure 3.14). We discovered that on average, the run time of DTTox can be saved by 30% with a 2% sacrifice in model performance. Further, the run time of DTTox can be cut in half with a 5% sacrifice in model performance. These statistics offer DTTox users some flexibility in balancing model efficiency and performance, especially when implementing the model on larger datasets.

3.4 Discussion

Biologically informed VNNs provide a solution to the dilemma posed by conventional supervised learning models: Whether to achieve good predictive performance or high model interpretability. Here, we have explored the implementation of VNNs for predicting and explaining compound response to toxicity assays. Compared to previous efforts, our DTTox model uniquely stands out in four aspects: First, the structure of DTTox can be customized for an outcome of interest according to the underlying biological processes, making the model flexible towards various toxicity outcomes. It also provides a molecular overview of each toxicity outcome regarding which pathway categories are relevant to the outcome. The molecular overview can help prioritize areas for future research in drug discovery as many toxicity outcomes we analyzed are linked to complex diseases. For instance, the androgen receptor, estrogen receptor, and progesterone receptor all exhibit aberrant activities in various cancer types and play a central role in the progression and metastasis of these types including breast cancer, prostate cancer, ovarian cancer. Both constitutive androstane receptor and pregnane X receptor can regulate drug-metabolizing enzymes and transporters, and thus play a critical role in resistance to cancer therapy and other adverse drug reactions. Second, trimming of network hierarchy can remove unrelated pathways from the network and significantly reduce the number of trainable parameters in VNNs, which in turn prevents overfitting. Through comparisons with well-established classification algorithms, we have demonstrated that DTTox is a highly efficient learning model with good predictive performance. For instance, DTTox achieved the same level of performance as a matched MLP with only three percent of the network parameters. Through shuffling analysis, we have demonstrated that DTTox benefits from the incorporation of Reactome pathway knowledge,

including protein-pathway annotations and child-parent pathway relationships. Shuffling child-parent pathway relationships (higher hierarchy) exhibits a moderate impact on model performance compared to shuffling protein-pathway annotations (lower hierarchy), as there are fewer alternative pathways to sample from in the higher hierarchy. It also implies undocumented relationships other than child-parent, such as crosstalk between pathways from different branches, may play a critical role in some toxicity outcomes. Future investigation should be conducted how to train a VNN to recognize these interactions. Third, the introduction of an early stopping criterion combined with a relatively small network size makes DTox a fast-learning model. Last, and most importantly, DTox advances an interpretation framework that identifies high-relevance VNN paths for explaining the toxicity outcome of compounds. The framework builds on top of layer-wise relevance propagation¹⁰³ and assigns a relevance score to each VNN path. The innovation of our framework resides in its ability to statistically assess the significance of each path, with an empirical p -value computed from permutation testing. With the help of existing experimental datasets, we have validated the mechanistic interpretation by our framework and demonstrated the biological significance of DTox. For instance, we showed that DTox was able to consistently identify the corresponding “ground truth” VNN path representing mechanisms of transcription activation by three nuclear receptors. We employed Mechanism of Assay screening data to show that DTox was able to differentiate distinctive mechanisms leading to HepG2 cytotoxicity. We employed drug-induced transcriptome profiling data to show that DTox was able to disproportionately identify VNN paths representing the cellular activities induced by aromatase inhibitors and pregnane X receptor agonists, implying its potential to detect mechanisms of action.

Besides the expected results, DTox also generated new mechanistic hypotheses along model interpretation, some of which are supported by previous studies. For instance, our framework suggested a potential role for p53 in the regulation of aromatase in MCF-7aro/ERE, a breast cancer cell line. It has been revealed that p53 can directly bind to proximal promoter PII in breast adipose stromal cells, which in turn inhibits aromatase expression¹⁰⁸. In another case, our

framework suggested three signaling pathways, including MAPK/ERK (i.e., MAPK1/MAPK3 Reactome pathway), Hippo, and TLR3/4, contribute to the HepG2 cytotoxicity of mifepristone. Accordingly, recent studies have pointed out the effect of mifepristone on ERK activation¹¹⁰, YAP (a core factor of Hippo) activation¹¹¹, and TLR4 regulation¹¹². Particularly, ERK activation by mifepristone can lead to cytotoxicity in uterine natural killer cells¹¹⁰ while YAP activation by mifepristone can induce hepatomegaly in mice¹¹¹. Two additional findings from our cytotoxicity analysis have been corroborated by previous studies: (i) The involvement of mitochondria in antigen presentation via ATP synthase and mitochondrial calcium uniporter¹¹⁷, and (ii) the disruption of TLR3/4 signaling in DILI^{118, 119}. In addition, some unexpected findings by DTox are worth further investigation, such as the role of immune response in HepG2 cytotoxicity, the role of interleukin 3/14 in regulation of aromatase, the role of caspase 3/7 in MHC class II antigen presentation, etc.

3.5 Acknowledgements

This work was supported by NIH grants P30 ES013508, R01 LM010098, R01 AG066833, and K99 LM013646.

CHAPTER 4: KNOWLEDGE GRAPH AIDS COMPREHENSIVE EXPLANATION OF DRUG TOXICITY

This chapter was originally submitted as: Hao, Yun, Romano, Joseph, D., and Moore, Jason, H., “*Knowledge graph aids comprehensive explanation of drug toxicity.*” doi: 10.1101/2022.10.07.511348

Contributions:

J.H.M. and Y.H. conceived the AIDTox project. J.H.M., Y.H., and J.D.R. designed the AIDTox model and data analysis workflow (J.D.R. developed the ComptoxAI databased and helped with the method section 4.2.2). Y.H. and J.D.R. performed the analysis (J.D.R helped with the gene feature analysis in 4.3.3 and 4.3.4). J.H.M., Y.H., and J.D.R. interpreted the results and wrote the paper.

4.1 Introduction

Coupled with millions of data points generated by large-scale toxicity testing^{17, 18}, various QSAR models have been proposed to predict *in vitro* and *in vivo* endpoints⁴⁷⁻⁵⁵. While some have achieved decent predictive performance, almost none can overcome the trade-off between accuracy and interpretability⁵⁹. Under state-of-the-art models, it remains challenging to explain the toxicity outcomes of a compound with cellular activities involving target proteins, specific pathways, and biological processes. This limitation has raised substantial concerns among experimental toxicologists, calling for prediction models that can provide insight into cellular mechanisms of toxicity.

Recent developments in visible neural networks (VNN) provide a solution to the issue^{96, 98, 99}. In contrast to black-box neural networks, connections within VNNs are guided by curated knowledge from pathway ontologies. Specifically, the fully connected structure is replaced by an interpretable ontological hierarchy that connects input gene features to output response via hidden pathway modules. In a previous study, we developed a VNN model—named DTTox—for predicting compound response to toxicity assays¹²¹. Importantly, DTTox advances an innovative interpretation framework that identifies network paths connecting genes and pathways for

explaining the toxicity outcome of compounds. We demonstrated that DTox can achieve the same level of predictive performance as state-of-the-art models with a significant improvement in interpretability, as the identified paths can be linked to well-established mechanisms of toxicity. However, one major limitation of DTox is that the input feature profile is derived from structure-based binding prediction models. While these models ensure the wide applicability of DTox, they are vulnerable to prediction errors and data scarcity, causing exclusion of certain genes from the feature space. To address this limitation, we used curated knowledge from the toxicology-focused graph knowledge base ComptoxAI¹²² to refine the input feature profile of DTox. ComptoxAI provides extensive profiling of chemical-gene connections across multiple gene categories, which can be used for feature profile construction. Hence, the resulting model—named AIDTox (ComptoxAI + DTox)—contains many novel gene features with active roles in cellular mechanisms of toxicity, including tubulin proteins that regulate apoptosis signaling, cytochrome P450 enzymes that are mainly responsible for drug metabolism, and transporters that participate in drug elimination. We believe AIDTox will facilitate the generation of new hypotheses for mechanistic investigation. Our code can be accessed openly at <https://github.com/yhao-compbio/AIDTox>.

4.2 Materials and Methods

4.2.1 Processing cell viability screening datasets for model training

The Tox21 datasets¹⁷ contain screening results indicating the response of *in vitro* toxicity assays to compounds of interest. We extracted active and inactive compounds from the screening results of each assay, then removed compounds with inconclusive or ambiguous results. We focused our analyses on two cytotoxicity assays: HEK293 and HepG2, for each of which at least 500 compounds are available for model training.

4.2.2 Extracting chemical-gene connections from ComptoxAI for model construction

ComptoxAI^{122, 123} is a comprehensive graph knowledge base that contains curated relationships between chemicals, genes, assays, and many other entities. It contains two types of

relationships linking compound and gene nodes: physical binding (with protein product) and expression-alteration (up-regulation/down-regulation). We extracted both types for chemicals present in the Tox21 datasets. We also constructed a hybrid type by combining the two types of relationships. We used the extracted relationships to assemble binary compound-gene matrices as input feature profiles.

To perform dimensionality reduction, we implemented a ReliefF-based method—namely, MultiSURF⁶⁷—for ranking genes by relevance to the assay outcome of interest. MultiSURF takes in a labeled feature dataset (assay outcome as the label), ranks all features based on differences among neighboring instances. Specifically, intraclass differences will contribute negatively to feature relevance while interclass differences will contribute positively. A benchmark study showed that MultiSURF outperformed other methods in detecting genotype-phenotype associations⁶⁷. We selected the top 100, 200, 300, 400, and 500 ranked genes of each dataset to proceed with the following analysis. Model selection by training loss was carried out to identify the optimal gene feature space.

4.2.3 Constructing VNN with selected gene features and Reactome pathway hierarchy

DTox is a VNN model embedded with the Reactome pathway hierarchy that comprises root biological processes, child-parent pathway relations, and gene-pathway annotations (downloaded in Aug 2019)⁷⁵. Each pathway is embedded as a module consisting of hidden neurons. AIDox inherits the basic structure of DTox while modifying input layers to consist of the gene features selected by MultiSURF. Guided by gene-pathway annotations, the gene features are connected to modules representing lowest-level pathways. These modules are then connected to deeper layers based on child-parent pathway relations until root biological processes are reached. Finally, the root biological processes are connected to an output layer representing the assay outcome. To trim the network's scale and prevent overfitting, we made the root biological process a hyperparameter of AIDTox. We selected four processes ('gene expression', 'immune system', 'metabolism', and 'signal transduction') and all possible

combinations among them for the tuning process (15 values in total). These four processes were selected due to their broad coverage and direct involvement in cellular mechanisms of toxicity.

AIDTox also inherits the hybrid loss function of DTox combining both root and auxiliary loss terms:

$$BCELoss(y_{root}, y) + \alpha \sum_p \beta_p BCELoss(y'_{p-auxi}, y) + \lambda \|W\|_2$$

The first term, namely root loss, denotes the binary cross entropy of final output y_{root} . The second term, namely auxiliary loss, denotes the binary cross entropy of the auxiliary scalar y'_{p-auxi} from each pathway module, with factor α ($= 0.5$) balancing the root and auxiliary terms, and factor β_p (computed as the inverse of pathway count within the corresponding hidden layer) normalizing terms across hidden layers. The third term denotes L_2 regularization with coefficient λ ($= 1e^{-4}$). The incorporation of auxiliary loss terms can prevent gradients from vanishing in the lower hierarchy and facilitates learning new patterns from individual pathways.

4.2.4 Learning optimal AIDTox model for cytotoxicity prediction and interpretation

AIDTox adopts the same training scheme as DTox, including a training/testing/validation split by ratio of 7:1:2, optimization using the Adam algorithm¹²⁴ with mini-batch size of 32, a learning rate of 0.001, early stopping criterion with “patience” of 20, and hyperparameter tuning by grid search. We evaluated the performance of the optimal AIDTox model on the held-out validation set, and compared it with three existing models: (i.) the optimal DTox model, (ii.) an optimal QSAR model based on a random forest classifier (derived from tuning six hyperparameters with 2800 combinations in total, listed in Table 3.1), and (iii.) an optimal QSAR model based on a gradient boosting classifier (derived from tuning five hyperparameters with 3000 combinations in total, listed in Table 3.1). For the two QSAR models (ii. and iii.), we adopted 166-bit binary MACCS fingerprints to quantify structural properties of compounds, which covers most of the interesting physicochemical features for drug discovery¹⁰¹. We also implemented DTox’s model interpretation framework to identify paths from the VNN that can explain compound cytotoxicity.

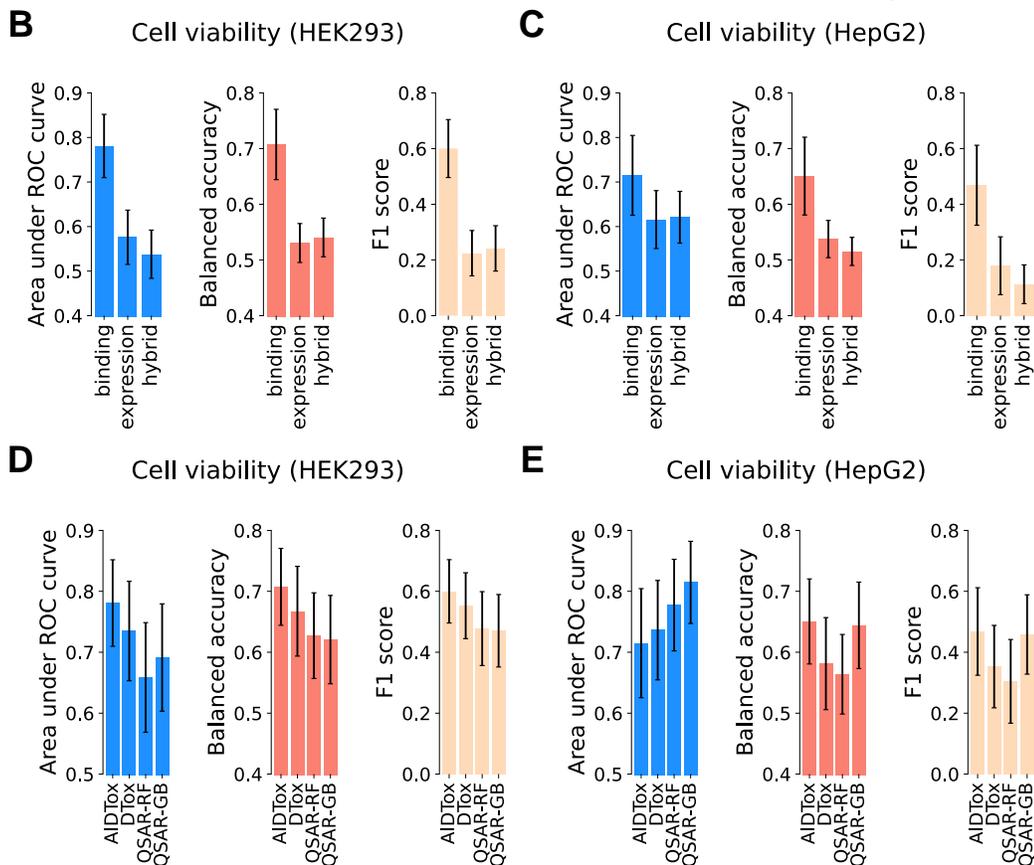
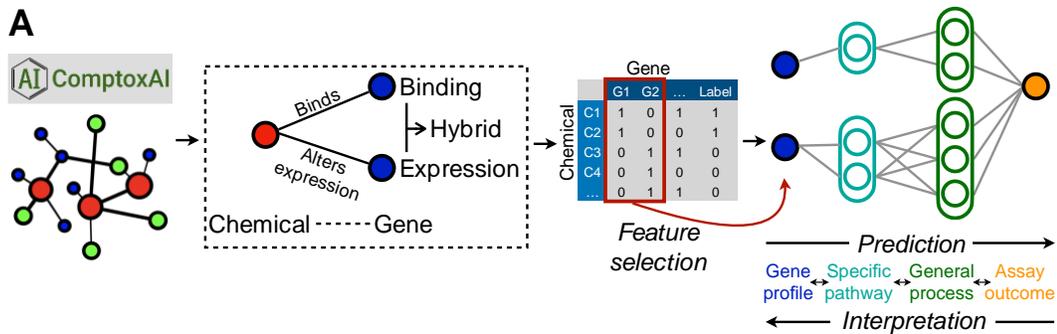


Figure 4.1 Incorporating curated chemical-gene connections into VNN for toxicity prediction with AIDTox

(A) Three types of chemical-gene connections (binding, expression, and hybrid) are extracted from ComptoxAI to construct the input feature profile of AIDTox. Feature selection is implemented to identify the top gene features predictive of the outcome of interest. The selected profile is fed into a VNN, whose structure is guided by Reactome pathway hierarchy. Specific pathways and general processes are coded as modules by hidden neurons. (B&C) Barplots showing the comparison of validation performance across three connection types in two cell viability datasets: HEK293 (B) and HepG2 (C). Performance is measured by three metrics: area under ROC curve, balanced accuracy, and F1 score, with error bar showing the 95% confidence interval. (D&E) Barplots showing the comparison of validation performance across four models in two cell viability datasets: HEK293 (D) and HepG2 (E). Three other models are considered: (i) our previous DTTox model with inferred target profile as input, (ii) QSAR model by random forest with chemical fingerprint as input, and (iii) QSAR model by gradient boosting with chemical fingerprint as input.

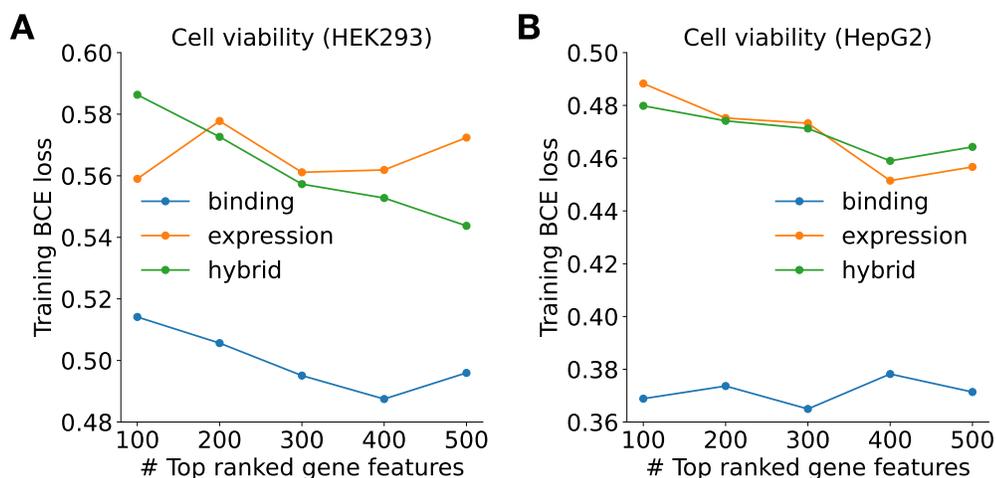


Figure 4.2 Relationship between training performance and the number of top predictive gene features

Line charts showing the relationship between training performance (y-axis) and the number of top predictive gene features (x-axis) in two datasets: HEK293 (**A**) and HepG2 (**B**) cell viability. Training performance is visualized by the binary cross entropy loss (BCE loss) for models derived from three types of chemical-gene connections: binding (blue line), expression (orange line), and hybrid (green line).

4.3 Results

4.3.1 AIDTox employs curated chemical-gene connections to construct VNN

AIDTox predicts and explains compound response to toxicity assays with a knowledge-guided VNN. The knowledge incorporated in AIDTox comprises chemical-gene connections from ComptoxAI, as well as gene-pathway annotations and child-parent pathway relationships from Reactome (Figure 4.1A). Connections within the VNN are constrained to gene-pathway connections (input to first hidden layer) and child-parent pathway relations (after first hidden layer). In this study, we focused on two cell viability assays measuring compound cytotoxicity in HEK293 and HepG2 cells. Both datasets contain 1,367 compounds with connections in ComptoxAI, including 432 cytotoxic and 935 non-cytotoxic compounds in HEK293 cells, and 293 cytotoxic and 1,074 non-cytotoxic compounds in HepG2 cells. Among these compounds, 617 (45%) are connected to 991 genes with physical binding evidence, 1,237 (90%) are connected to 8,723 genes with expression-alteration evidence, and 1,367 (100%) are connected to 8,735 genes with both types of evidence. In all three cases, the number of gene features is much greater than the number of compound samples. To prevent overfitting, we reduced feature dimensions by selecting the top 100-500 predictive genes during model construction. In general,

as more predictive genes were incorporated, we observed an improvement in the resulting model performance (decline in training loss) until 400 (Figure 4.2). Therefore, we proceeded with models built using the top 400 predictive genes.

4.3.2 Chemical-gene binding connections result in the best performing models

Using held-out validation sets of HEK293 and HepG2 cell viability data, we first compared the performance of models derived from three distinct connection types. In both datasets, models derived from binding connections significantly outperform the other types (Figure 4.1B&C), as we observed no overlaps between the 95% confidence intervals of compared metrics (AUROC, balanced accuracy, and F1-score) except for AUROC on the HepG2 viability dataset. We then compared the performance of AIDTox models derived from binding connections against three other classification algorithms, including our previous DTox model (derived from predicted compound-target interactions) and two QSAR models (derived from quantified structural properties). In both datasets, the 95% confidence intervals of performance metrics are highly overlapped among compared algorithms (Figure 4.1D&E). Therefore, AIDTox achieved the same level of predictive performance as these well-established classification algorithms.

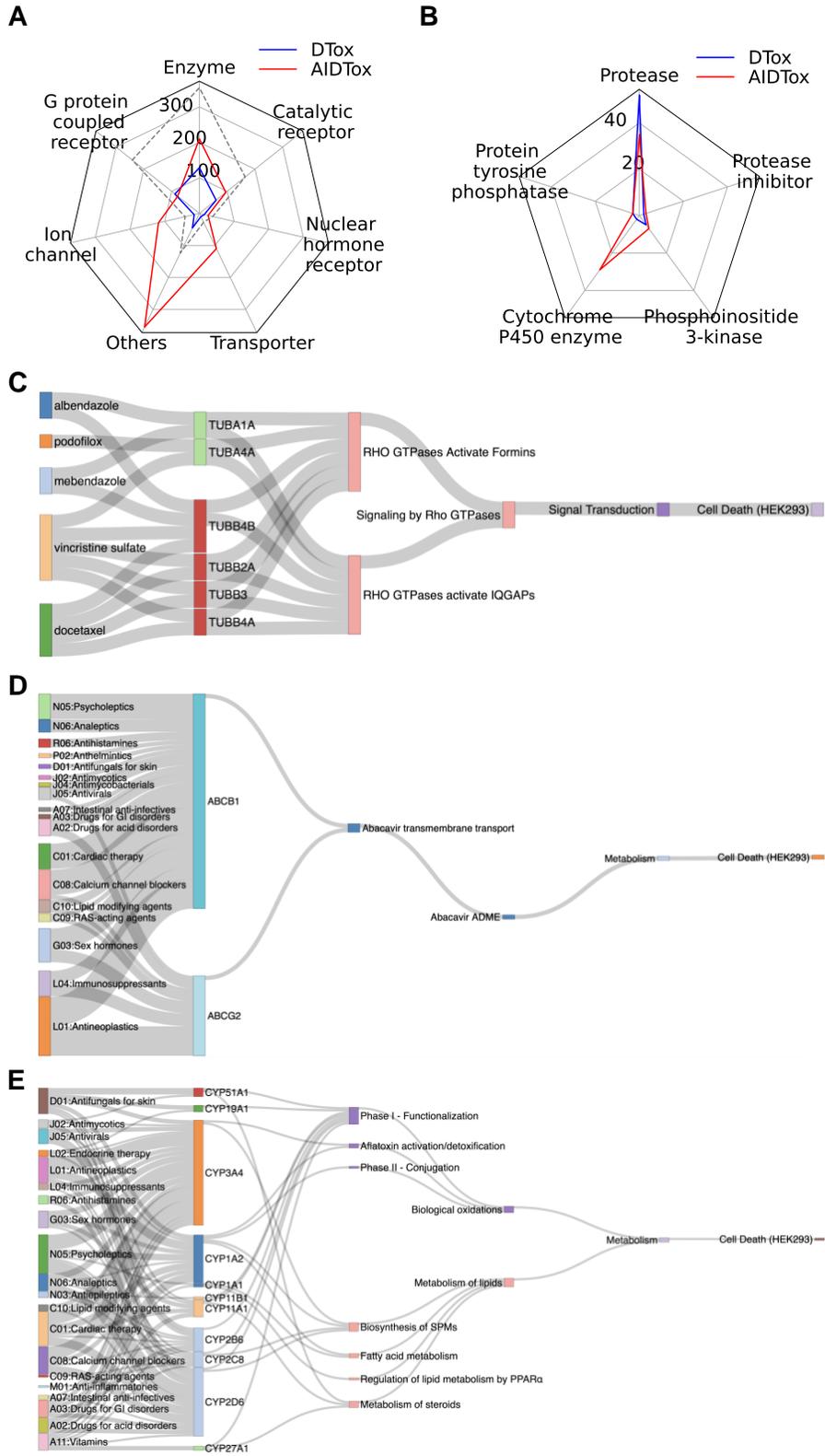


Figure 4.3 Comprehensive explanation of HEK293 cytotoxicity with new features in AIDTox

(A) Radar plot showing the comparison of gene category distributions among the features of DTox (blue solid line) and AIDTox (red solid line). The grey dashed line shows the hypothetical of a proportionate increase from DTox to AIDTox. (B) Radar plot showing the comparison of enzyme subcategory distributions among the features of DTox (blue solid line) and AIDTox (red solid line). (C) Sankey diagram showing the AIDTox explanation of HEK293 cytotoxicity for drugs targeting tubulin proteins. The paths (connecting drugs to HEK293 cell death) shown in the diagram are identified from the full network of VNN model by the AIDTox interpretation framework. Connections in the VNN are informed by ComptoxAI (chemical-gene) and Reactome (gene-pathway, child-parent pathway). Tubulin proteins are grouped and colored by the family. (D) Sankey diagram showing the AIDTox explanation of HEK293 cytotoxicity via ATP-binding cassette transporters (similar to C). Drugs are grouped and colored by the ATC subclass (first three digits). (E) Sankey diagram showing the AIDTox explanation of HEK293 cytotoxicity via cytochrome P450 enzymes (similar to C). Drugs are grouped and colored by the ATC subclass (first three digits). Cytochrome P450 enzymes are grouped and colored by the family. Metabolic pathways are grouped and colored by the general metabolic process they belong to (“Biological oxidations” or “Metabolism of lipids”).

4.3.3 AIDTox models benefit from a comprehensive gene feature space

A fundamental advantage of AIDTox over DTox comes from its enlarged gene feature space, providing an extensive profiling of the cellular activities of compounds. Overall, there is an increase from 361 genes in DTox to 991 genes in AIDTox, a 2.7-fold increase. Comparing the composition of gene features, we discovered that the increase in AIDTox is mainly driven by three target categories: ion channels, transporters, and others (Figure 4.3A). One example of the “others” category is the tubulin protein superfamily (Figure 4.3C), which polymerize into microtubules, a primary component of cytoskeleton involved in cell division. Due to the lack of binding data, tubulin proteins were not present in the DTox model. In ComptoxAI, the two families of tubulin proteins (α and β) are connected to antiparasitics (albendazole, and mebendazole) and antineoplastics (podofilox, vincristine sulfate, and docetaxel). Our AIDTox model further connects these drugs to HEK293 cytotoxicity via Rho GTPase signaling by effectors such as formins and IQGAPs, which have been shown to regulate apoptosis in HEK293 cells¹²⁵.

4.3.4 New features in AIDTox are essential in drug metabolism and elimination processes

We observed a disproportionate 11-fold increase (from 10 to 110) among transporters in the gene feature space of AIDTox. For instance, AIDTox connects drugs within nine ATC classes (18 subclasses) to HEK293 cytotoxicity via two new features: transporters *ABCB1* and *ABCG2* (Figure 4.3D). These two ATP-binding cassette transporters are mainly responsible for pumping

drugs out of the cell, thus they play a critical role in drug elimination. Therefore, inhibition of transporter-mediated elimination by drugs may prolong their cytotoxic effect on HEK293 cells¹²⁶.

While we did not observe a disproportionate increase among the general enzyme category, we did observe a 14.5-fold increase (from 2 to 29) among one subcategory: the cytochrome P450 enzymes (CYPs; Figure 4.3B). For instance, AIDTox connects drugs within nine ATC classes (20 subclasses) to HEK293 cytotoxicity via 20 CYPs of seven families (Figure 4.3E). These CYPs are the major enzymes involved in drug metabolism as they participate in metabolic pathways such as “biological oxidation” and “metabolism of lipids”. Meanwhile, CYP-mediated metabolic processes are the main source of reactive oxygen species, which can lead to cellular oxidative stress and trigger apoptosis/necrosis¹²⁷.

4.4 Discussion

Rich knowledge in ComptoxAI provides an accurate and extensive profiling of chemical-gene connections. Here, we have explored the incorporation of these connections into knowledge-guided deep learning models for predicting and explaining compound cytotoxicity. We considered three types of connections for the task: physical binding, expression-alteration, and a hybrid type. Models derived from binding connections exhibit the best predictive performance, since physical binding is stronger evidence of direct interaction compared to expression alteration. In contrast to our previous work DTTox, the new AIDTox model employs curated knowledge for generating input feature profile, thus is not prone to errors from binding prediction models. In addition, AIDTox is not restrained by the availability of compound-target binding data. The feature space of AIDTox comprises many genes with insufficient binding data, including ion channels, transporters, and CYP enzymes. These categories exhibit central roles in drug metabolism and elimination processes. Accordingly, they become an asset for prediction and explanation of complex toxicity outcomes, which may be triggered by multiple cellular mechanisms. For instance, AIDTox was able to connect dasatinib, a leukemia drug, to HEK293 cytotoxicity via multiple aspects of drug activity, including MAPK14-mediated regulation of apoptosis, CYP1A2/CYP3A4-mediated drug metabolism, and ABCB1/ABCG2-mediated drug elimination (Figure 4.4). It is also worth noticing

the high interpretability of AIDTox is achieved without loss of accuracy. Therefore, we anticipate AIDTox to be applied in both prioritizing compounds for safety testing and generating new hypothesis for mechanistic investigation.

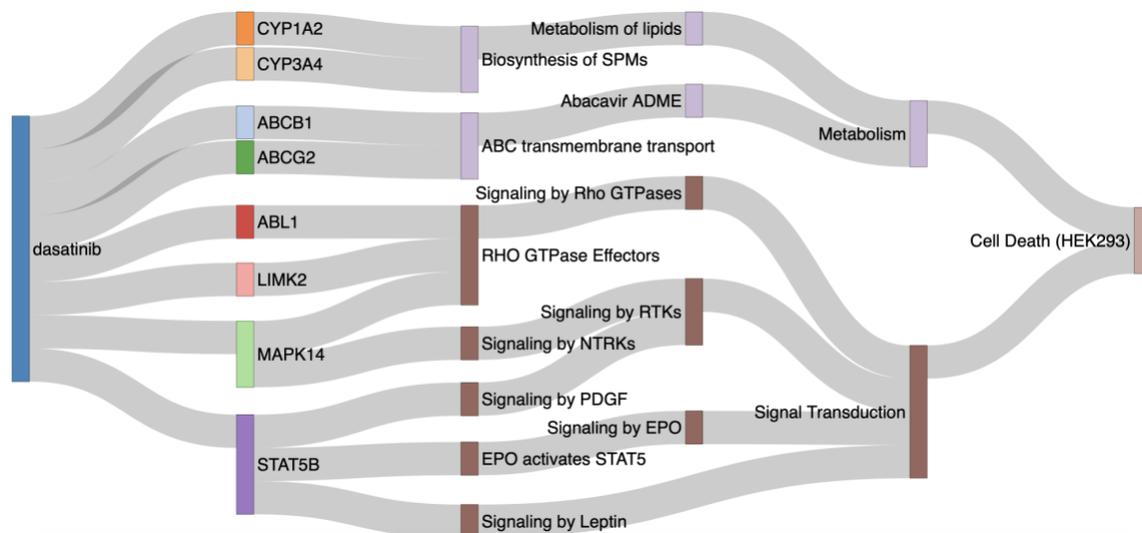


Figure 4.4 AIDTox explanation for HEK293 cytotoxicity of dasatinib

Sankey diagram showing the AIDTox explanation of HEK293 cytotoxicity for dasatinib, a drug used for leukemia treatment. The paths (connecting dasatinib to HEK293 cell death) shown in the diagram are identified from the full network of VNN model by the AIDTox interpretation framework. Connections in the VNN are informed by ComptoxAI (chemical-gene) and Reactome (gene-pathway, child-parent pathway). Pathways are grouped and colored by the general process they belong to (“Metabolism” or “Signal Transduction”).

Despite these highlights, AIDTox did not significantly outperform existing methods in cytotoxicity prediction. This can be attributed to the sharp decline in sample size, as the profiling of chemical-gene connections by ComptoxAI only applies to well-studied compounds. A smaller sample size in turn leads to wider confidence intervals of model performance metrics. Consequently, even though we observed a moderate increase in the performance metrics, the increase is well within the 95% confidence interval. In the future, we expect the performance of AIDTox to be enhanced after chemical-gene connections become available for more compounds. We also acknowledge the recent development of graph neural network-based link prediction algorithms, which can fill in missing connections for under-studied compounds. Such algorithms may help us increase the sample size for AIDTox training and enhance model performance. In

addition to increasing sample size, we think the incorporation of context-specific knowledge, such as cell line-specific chemical-gene connections, may further enhance the performance of AIDTox.

4.5 Acknowledgements

This work was supported by NIH grants P30 ES013508, R01 LM010098, R01 AG066833, and K99 LM013646.

CHAPTER 5: DISCUSSION AND FUTURE DIRECTIONS

In this thesis work, we aimed to improve the interpretation of toxicity prediction, addressing a critical drawback of existing *in silico* models. We developed a series of interpretable models: TargetTox, DTox, and AIDTox, to connect chemicals to their toxicity targets and pathways. TargetTox is a feature selection pipeline that can identify predictive targets associated with drug toxicity. DTox is a knowledge-guided deep learning model that can identify cellular events for explaining toxicity outcomes of individual compounds. AIDTox is an enhanced version of DTox that incorporates connections from knowledge graph for achieving comprehensive explanation of drug toxicity. We demonstrated that the identified targets and pathways can help explain cellular mechanisms underlying structure-toxicity associations. In this chapter, we will discuss the highlights and limitations of our models, as well as the future directions of *in silico* toxicity assessment.

5.1 Summary and highlights of TargetTox

In our first model—TargetTox, we proposed a prediction model that relates structure properties to toxicity outcomes via target profile of drugs. Compared to conventional QSAR models, target-based prediction appears to be a better alternative as drugs causing same adverse events often share targets, making it easier for classification algorithms to detect predictive features that can separate positive instances from negative controls. In TargetTox, we incorporated a novel feature selection pipeline to resolve data dimensionality issues regarding small sample space versus large feature space. We implemented TargetTox to identify predictive descriptors for target binding, as well as predictive targets for adverse events. In both tasks, we were able to approximate the performance by all features with just the identified predictive features. More importantly, without sacrificing model accuracy, we were also able to significantly improve model interpretability by linking the predictive targets to cellular mechanisms of toxicity. For instance, the predictive targets tend to be differentially expressed in the tissue of toxicity. They are also enriched for key toxicity-related Gene Ontology terms, adverse outcome pathways, and markers/therapeutics of the matched disease category. Combined together, our results

highlighted the critical roles of predictive targets in adverse drug reactions. These targets often exhibit specific functions in the tissue of toxicity. Perturbation of the functions by compounds could lead to aberrant activities of cellular pathways, triggering adverse drug reactions.

5.2 Limitations of TargetTox and future directions

Despite these highlights of TargetTox, it did not significantly outperform structure-based QSAR models in adverse event prediction. The AUROC of TargetTox models ranges from 0.5 to 0.7, which is mostly in line with previous efforts predicting adverse events, including conventional QSAR models and hybrid models (combining structural features with other feature types). There are multiple limiting factors contributing to the unsatisfying performance of TargetTox in adverse event prediction. In the following paragraphs, we will make a detailed discussion of each factor and propose potential directions for improvement.

First and foremost, TargetTox is a data-driven model in that the target profile used for adverse event prediction is derived from ligand-based binding prediction models. This setting ensures the general applicability of TargetTox, as binding prediction models only require structural features of compounds as input. However, the training of binding prediction models largely relies on the availability of compound-target binding affinity data. A target protein will be absent from the feature profile if there is a lack of binding affinity data for model training, or the binding outcome cannot be accurately predicted by structural features. In TargetTox, ~400 target proteins were included in the feature profile for adverse event prediction, which only accounts for a small proportion of the proteins in human druggable genome (4,000-5,000 proteins by estimation). The incomplete profile of target proteins may negatively affect the predictive performance and interpretability of TargetTox if the certain targets play a critical role in adverse drug reaction. While ligand-based models may benefit from release of more compound-target binding affinity data, using them for profiling of whole human druggable genome may be a far fetch considering the technical difficulties of data acquisition for certain proteins (i.e., those with low abundance in human body). One potential solution is to adopt interaction-based binding prediction models for target profiling¹²⁸⁻¹³⁰. These models take into account both compound

structure and protein sequence and learn a generic model from millions of compound-target pairs, thus can be applied to any protein in human druggable genome. However, it can be quite challenging to apply a single interaction-based binding prediction model to all the druggable proteins, as the binding mechanism for different protein classes can vary drastically. Such challenge hinders the immediate adoption of interaction-based binding prediction in target profiling. We hope future studies can provide a comprehensive evaluation of interaction-based binding predictions and shed light on their applicability among distinct protein classes.

Second, the target profile used for adverse event prediction is binary in the current form of TargetTox (i.e. whether a compound binds to the target or not). Substantial evidence has pointed to the importance of binding direction in adverse drug reactions (i.e. activation by agonists or inhibition by antagonists). For instance, as we illustrated in Section 1.2, cardiotoxicity of tyrosine kinase inhibitors is linked to the inhibition of EGFR signaling in normal tissues while cardiotoxicity of terfenadine is linked to inhibition of cardiac ion channels (hERG). The incorporation of binding direction will turn the binary classification by TargetTox into three-class prediction problem. Currently, the direction of binding can be modeled for a small number of target proteins using data source such as Tox21, which contains screening results of agonist/antagonist assay. However, due to a lack of directed compound-target binding data, the modeling cannot be accomplished for most proteins in human druggable genome. The issue can only be resolved when more relevant data becomes available.

Lastly, TargetTox models the occurrence of adverse drug reactions entirely on compound-target interactions. However, as we illustrated in Section 1.2, drug toxicity can also be caused by oxidation-reduction reactions of toxic metabolites after biological transformation, hypersensitivity and immunological reactions, etc. Therefore, adverse drug reactions may not be fully explained by the variation in target binding, contributing to the relatively poor predictive performance for some adverse events. One potential solution is to construct hybrid toxicity prediction models by combining target binding profile with other feature types, including drug-induced gene expression signatures⁵⁷, extracted features characterizing dose-response curves¹³¹, physiological

parameters generated from pharmacodynamic and pharmacokinetic modeling¹³², etc. These feature types reflect the absorption, distribution, metabolism, and excretion processes of drugs, as well as drug-induced perturbation on a transcriptome scale, thus may account for the cellular mechanism of toxicity unrelated to binding of specific targets. However, it is worth mentioning that these feature types often cannot be directly computed or inferred from compound structure. To obtain feature values, *in vitro* experiments have to be performed, limiting the applicability of such integrated models.

Apart from the unsatisfying performance in adverse event prediction, the uncertainty of TargetTox predictions remains high, as we observed wide confidence intervals for performance metrics. This can be attributed to the imbalanced ratio of positive to negative samples (the number of negative samples is often much greater than the number of positive samples) as we observed a strong correlation between the ratio and confidence interval width. Class balancing techniques such as upsampling minority class and downsampling majority class can be adopted to address the issue.

5.3 Summary and highlights of DTox

In our second model—DTox, we proposed a biologically informed VNN that overcomes the accuracy-interpretability trade-off faced by previous toxicity prediction models. In contrast to conventional deep learning models coupled with *post hoc* interpretation techniques, the structure of DTox is guided by extensive knowledge from pathway ontology connecting structural properties to toxicity endpoints via target proteins, specific pathways, and biological processes, making the model interpretable by design. DTox is innovative because it represents a departure from the status quo by incorporating prior biological knowledge into model construction. In doing so, model interpretation is linked to model training. In addition, DTox interpretation is not affected by model complexity, whereas *post hoc* interpretation often becomes challenging for complex neural networks.

As we illustrated in Section 3.4, compared to previous efforts of VNN development, our DTox model uniquely stands out in four aspects: customized structure to increase flexibility, reduced

number of trainable parameters to prevent overfitting, early stopping criterion to speed up training, and innovative interpretability framework to explain model predictions. We implemented DTox to predict and explain compound response to 15 *in vitro* toxicity endpoints including nuclear receptor signaling, stress response, cytotoxicity, and developmental toxicity. We demonstrated that DTox is a highly efficient learning model with the same level of predictive performance as well-established classification algorithms. In particular, DTox achieved this with only three percent of the network parameters of a matched MLP model. We also demonstrated the utility of DTox interpretation framework in facilitating *in silico* mechanistic investigation, therefore showcase its biological significance. For instance, in three assays measuring nuclear receptor activation, DTox consistently identified the “ground truth” VNN path that represents well-established mechanism of transcription activation by the corresponding receptor. In another two assays (aromatase inhibitor and pregnane X receptor agonist), DTox disproportionately identified the VNN paths that represent the cellular activities leading to the corresponding outcome. In the HepG2 viability assay, DTox was able to differentiate distinctive mechanisms leading to cell death by giving higher relevance to the relevant pathway module.

5.4 Limitations of DTox and future directions

Despite the highlights mentioned above, DTox in its current form bears some limitations from both technical and methodological perspectives. In terms of technical limitations, as with all deep learning models, DTox requires a time-consuming hyperparameter tuning process before an optimal model can be reached. And as we illustrated in Figure 3-4A, an optimal setting can greatly improve the predictive performance of DTox. Since it takes an average of 19.6 hours to train a DTox model on Tox21 datasets, the hyperparameter tuning process may take days or even weeks to complete on a single CPU. However, the issue can be resolved with implementation of graphics processing unit (GPU) computing.

In terms of methodological limitations, DTox did not significantly outperform other well-established classification algorithms, as most differences are within the 95% confidence interval of performance metrics. In the interpretation analyses, DTox was not able to identify the ground

truth path for a particular assay: Androgen receptor antagonist. It also failed to identify more differentially expressed paths in general for two assays: Mitochondria toxicity and HepG2 cell viability. Since DTox models the *in vitro* toxicity endpoints entirely on compound-target interactions, these results imply that additional factors unaccounted for in DTox may also play a critical role leading to toxicity, including oxidation-reduction reactions of toxic metabolites after biological transformation, hypersensitivity and immunological reactions, similar to the case of TargetTox. In this study, we emphasized the applicability of DTox such that it can be adopted to study any compounds with or without additional profiling information. Therefore, we limited the model input to the 2D structural representation of compounds. In the future, we expect the performance of DTox to be enhanced after incorporation of additional profiles, such as drug-induced gene expression signatures, extracted features characterizing dose-response curves, physiological parameters generated from pharmacodynamic and pharmacokinetic modeling, as we discussed in Section 5.2. Among these feature types, drug-induced gene expression signatures can be directly connected to the hidden layers of a VNN based on gene-pathway annotations, while quantified dose-response features and physiological parameters cannot. The latter two feature types can be embedded in a conventional artificial neural network (ANN) in parallel with the VNN model. Then the two branches of the hybrid model, the VNN embedding target profile, and the ANN embedding additional features can be combined in a single layer of neurons⁹⁷.

Another limitation of DTox concerns the dose-dependent effect of drug toxicity. Currently, DTox is a dose-naïve model that makes toxicity predictions purely based on the input target binding profile, which is derived from structural properties of compounds. However, there is substantial evidence pointing the critical role of dose on drug toxicity¹³³. For most endpoints, toxicity often occurs at doses that exceed the therapeutic efficacy of drugs¹³⁴. Therefore, a toxicity prediction model that takes dose into consideration is urgently needed. While it may be challenging to directly incorporate dose as a feature in DTox, the hybrid models proposed above can be viewed as indirect incorporation. That is because the drug-induced gene expression

signatures, quantified dose-response features, and physiological parameters can reflect the effect of drug dose on cellular activities.

In addition to new feature types, we think the incorporation of context-specific knowledge may further enhance the performance of DTTox. As we demonstrated in Figure 3.5 and discussed in Section 3.4, undocumented interactions between pathways may play a critical role in some toxicity outcomes. These interactions can be specific to the outcome of interest, thus are often missing from generic resources such as Reactome or the Gene Ontology. In the past few years, many context-specific gene networks were developed, including GIANT tissue-specific networks^{135, 136}, YETI tissue-specific networks¹³⁷, etc. These networks cover a diverse range of human tissues and diseases, and can inform us the context-specific connections between pathways, providing a solution to the incorporation of context-specific knowledge in DTTox. Another way to address the limitation is to incorporate stochastic connections between pathways of distinct branches during DTTox training, forcing the model to learn meaningful new interactions that aid the prediction of each outcome.

Finally, in the DTTox project, we employed experimental datasets from mechanism of action screening and drug-induced transcriptome profiling to validate the interpretation framework. In addition, DTTox interpretation can be validated by other types of experiments in the future. For instance, knockdown and overexpression experiments can be performed to evaluate the inferred causality between toxicity phenotypes and target proteins/pathways. For toxicity outcomes that can be linked to clinical phenotypes in patients, observation data from electronic health records can be employed to perform survival analysis, evaluating the inferred causality between the phenotype and lab measurements (e.g., enzyme level, cell count) that can inform the activities of certain proteins/pathways.

5.5 Summary and highlights of AIDTox

In addition to the limitations discussed above, another limitation of DTTox is that the input feature profile is derived from structure-based binding prediction models. While these models ensure the wide applicability of DTTox, they are vulnerable to prediction errors and data scarcity,

causing exclusion of certain genes from the feature space. In our third model—AIDTox, we addressed this limitation by using curated knowledge from the toxicology-focused graph knowledge base ComptoxAI to refine the input feature profile of DTox. ComptoxAI provides extensive profiling of chemical-gene connections across multiple gene categories. As a result, AIDTox contains many novel gene features with active roles in cellular mechanisms of toxicity, including tubulin proteins that regulate apoptosis signaling, cytochrome P450 enzymes that are mainly responsible for drug metabolism, and transporters that participate in drug elimination. These categories exhibit central roles in drug metabolism and elimination processes. We demonstrated the utility of AIDTox in predicting and explaining complex toxicity outcomes such as cytotoxicity, which are triggered by multiple cellular mechanisms. We also demonstrated that the high interpretability of AIDTox is achieved without loss of accuracy. While the applicability of AIDTox is not as wide as DTox at the moment (AIDTox is only applicable to one-tenth of the drugs compared to DTox), we expect it be enhanced after chemical-gene connections become available for more compounds, or graph neural network-based link prediction algorithms are implemented to fill in missing connections for under-studied compounds.

5.6 Graph neural network can improve the performance of toxicity prediction

This section was extracted from the paper originally published as: Romano, Joseph, D.*, Hao, Yun*, and Moore, Jason, H., “*Improving QSAR Modeling for Predictive Toxicology using Publicly Aggregated Semantic Graph Data and Graph Neural Networks.*” In PACIFIC SYMPOSIUM ON BIOCOMPUTING. 2022 pp:187-198. doi: 10.1142/9789811250477_0018

Contributions:

J.H.M., J.D.R., and Y.H. conceived the QSAR-GNN project. J.H.M., J.D.R., and Y.H. designed the model and data analysis workflow. J.D.R., and Y.H. performed the analysis (J.D.R. led the analysis in model construction and model training, Y.H. performed model comparison and result visualization). J.H.M., J.D.R., and Y.H. interpreted the results and wrote the paper.

Another major limitation of the three models proposed (TargetTox, DTox, and AIDTox) in this thesis work is concerned with their predictive performance: We were only able to achieve the same level of performance as conventional structural-based QSAR models by well-established classification algorithms such as random forest, gradient boosting, artificial neural network, etc. While the ultimate goal of this thesis is to improve the interpretability of toxicity prediction with biological knowledge, we would like to explore whether the incorporation of biological knowledge can improve the accuracy of toxicity prediction as well. The resulting model may serve as an alternative for users who value model accuracy more than interpretability. To this end, we again used ComptoxAI, which was introduced in AIDTox, as our knowledge source. ComptoxAI includes a large graph database – implemented as a knowledge graph – containing many entity and relationship types that pertain to translational mechanisms of toxicity, all of which are sourced from third-party public databases (including PubChem, Drugbank, the US EPA's Computational Toxicology Dashboard, NCBI Gene, and many others). To improve model accuracy, we augmented the traditional QSAR approach with multimodal graph data aggregated from ComptoxAI and analyzing those data in the context of a heterogeneous graph convolutional neural network (GCN) model. GCNs are a relatively new class of models based on artificial neural networks that have performed incredibly well on many prediction tasks involving densely connected biomedical data, including drug-drug interactions, protein function, and medical term semantic type prediction among others. Therefore, we first extracted the subgraph from ComptoxAI's graph database defined as all nodes representing chemicals, genes, and toxicological assays, as well as the complete set of edges linking nodes of those types. We then evaluated the model on 52 assays and their accompanying chemical screening data from the Tox21 dataset and compared its performance to two rigorously defined traditional QSAR models consisting of random forest and gradient boosting classifiers.

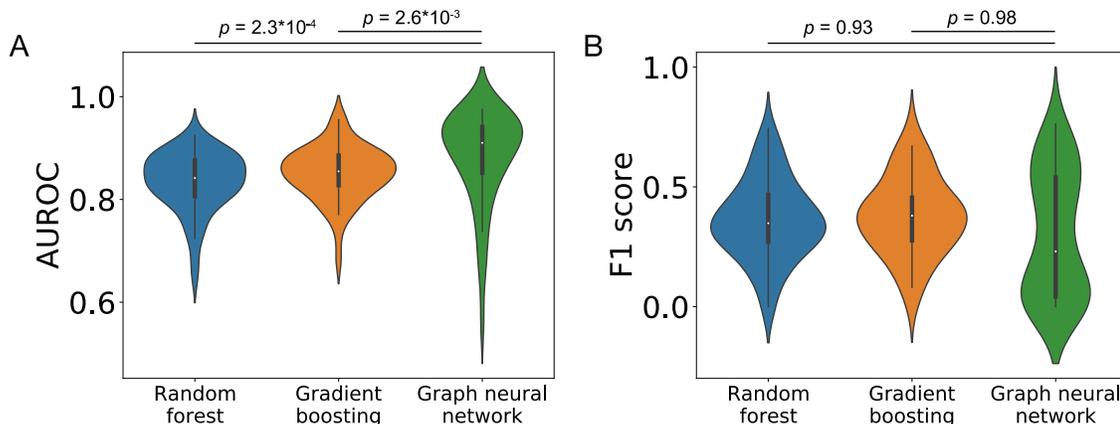


Figure 5.1 Overall performance metrics of the 3 QSAR model types on each of the Tox21 assays (A) AUROC and (B) F1 score. The mean AUROC is significantly higher for the GNN model than for either of the two baseline models. Differences in F1 scores are not statistically significant. The GNN achieves poor F1 scores on assays with relatively few (e.g., < 100) “active” annotations in Tox21, which is consistent with known performance of neural networks on data with sparse labels. p-values correspond to Wilcoxon signed-rank tests on means, with a significance level of 0.05.

We found that the GNN model significantly outperforms both the random forest ($P = 2.3e^{-4}$, Wilcoxon signed-rank test) and gradient boosting ($P = 2.6e^{-3}$) models in terms of AUROC (Figure 5.1A), with a mean AUROC of 0.883 (compared to 0.834 for random forest and 0.851 for gradient boosting). This is robust evidence that the GNN model tends to substantially outperform conventional QSAR models.

To better understand how the GNN model outperforms the random forest and gradient boosting models, we performed an ablation analysis on the two previously mentioned assays—pregnane X agonism and HepG2 cell viability. For both of the assays, we re-trained the model after removing specific components from the GNN: i) All assay nodes, ii) All gene nodes, iii) MACCS fingerprints for chemical nodes (replacing them with dummy variables so the structure of the network would remain the same).

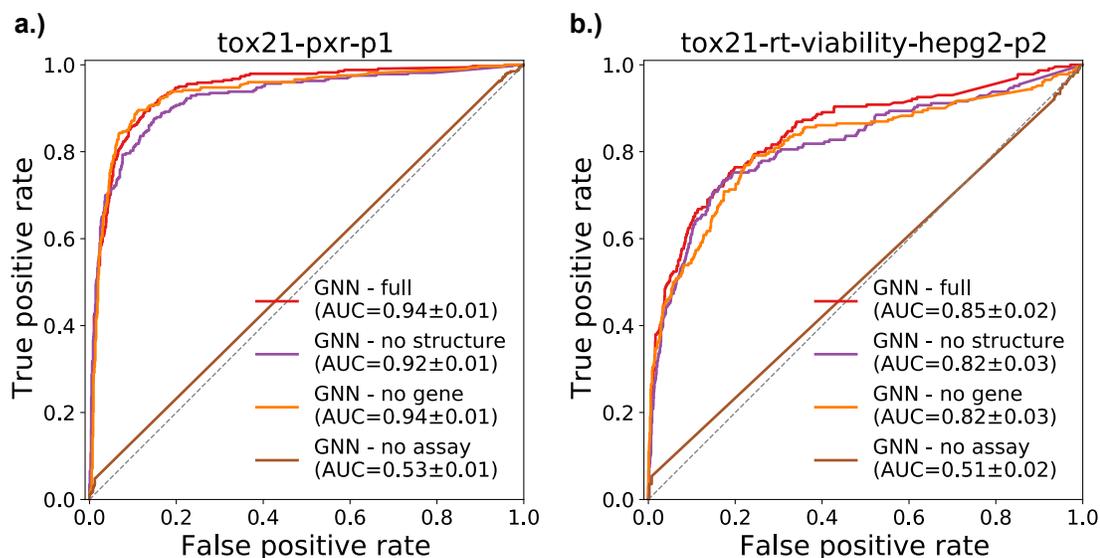


Figure 5.2 Receiver Operator Characteristic (ROC) curves for two selected Tox21 assays using different configurations of the GNN model

'GNN - full' is the complete model. 'GNN - no structure' omits the MACCS chemical descriptors and replaces them with node embeddings. 'GNN - no gene' omits gene nodes and their incident edges. 'GNN - no assay' removes all assay nodes and incident edges, so predictions are made solely using chemicals, genes, the remaining edges, and the MACCS fingerprints as chemical node features. AUC values include 95% confidence intervals.

As shown in Figure 5.2, for both assays, the full GNN model performed best, although only modestly better (in terms of AUROC) than the versions without MACCS fingerprints or gene nodes. However, the performance of the GNN drops substantially—barely better than guessing labels at random (i.e., AUROC = 0.5)—when assay nodes are removed from the graph. In other words, much of the inferential capacity of the GNN models is conferred by chemicals' connections to assays other than the one for which activity is being predicted. Similarly, MACCS fingerprints are not—on their own—enough for the GNN to attain equal performance to the baseline QSAR models, which only use MACCS fingerprints as predictive features. Therefore, although the GNN achieves significantly better performance than the two baseline models, it is only able to do so with the added context of network relationships between chemicals, assays, and (to a lesser degree) genes.

In summary, we introduced a novel GNN-based approach to QSAR modeling for toxicity prediction and evaluate it on data from 52 assays to show that it significantly outperforms existing methods. This work demonstrates that the incorporation of biological knowledge can improve the

accuracy of toxicity prediction, and our GNN model may serve as an alternative for users who value model accuracy more than interpretability. GNNs comprise an incredibly active emerging topic within artificial intelligence research, and as one of the first GNN applications in computational toxicology we hope that our results serve as a 'jumping off point' for a vast body of similar work. We plan to evaluate graph attention networks, new data modalities, and network regularization techniques in the near future.

5.7 Concluding remarks

In this thesis work, our goal is to improve the interpretability of toxicity prediction. To achieve the goal, we developed a series of interpretable models that connect drugs to their toxicity targets and pathways. We employed various experimental datasets to validate the mechanistic interpretation by our models and demonstrate their biological significance. Our models exhibit broad applicability in both drug discovery and chemical risk assessment, as no prior knowledge about compounds of interest is required other than the 2D structural representation.

In the future, we anticipate the application of our work in two distinct directions. The first direction is concerned with efficacy or toxicity prediction for virtual screening. As with what we have accomplished in the screening of ~700,000 DSSTox compounds for cytotoxicity, our prediction models can quickly go through large-scale chemical libraries and prioritize compounds for further experimental testing. The second direction is concerned with outcome explanation for generating new hypotheses. As we have shown throughout the study, our interpretation framework may detect new mechanisms of action for compounds, uncover cellular mechanism for outcomes of interest, and identify new therapeutic targets for diseases.

Our work substantially enhances the interpretability of toxicity prediction. It can provide more accurate and informative predictions of toxicity using publicly-available knowledge and data science. More importantly, it opens the door for using computational tools to conduct mechanistic investigation. As our developed models bridge together deep learning and molecular toxicology, users can gain a better understanding of the molecular basis of toxicity phenotypes. Due to these

highlights, we expect our work can reduce the burden of in vitro/in vivo testing, and facilitate chemical risk assessment and drug development.

BIBLIOGRAPHY

1. Kola, I. and J. Landis (2004). Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 3, 711-5.
2. Sacks, L.V., H.H. Shamsuddin, Y.I. Yasinskaya, K. Bouri, M.L. Lanthier, and R.E. Sherman (2014). Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000-2012. *JAMA* 311, 378-84.
3. Hay, M., D.W. Thomas, J.L. Craighead, C. Economides, and J. Rosenthal (2014). Clinical development success rates for investigational drugs. *Nat Biotechnol* 32, 40-51.
4. Scheiber, J., B. Chen, M. Milik, S.C.K. Sukuru, A. Bender, D. Mikhailov, S. Whitebread, J. Hamon, K. Azzaoui, and L. Urban (2009). Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *Journal of chemical information and modeling* 49, 308-317.
5. Liebler, D.C. and F.P. Guengerich (2005). Elucidating mechanisms of drug-induced toxicity. *Nature reviews Drug discovery* 4, 410-420.
6. Di, L. (2014). The role of drug metabolizing enzymes in clearance. *Expert opinion on drug metabolism & toxicology* 10, 379-393.
7. Deavall, D.G., E.A. Martin, J.M. Horner, and R. Roberts (2012). Drug-induced oxidative stress and toxicity. *Journal of toxicology* 2012.
8. Pichler, W.J., J. Adam, B. Daubner, T. Gentinetta, M. Keller, and D. Yerly (2010). Drug hypersensitivity reactions: pathomechanism and clinical symptoms. *Medical Clinics* 94, 645-664.
9. Segaert, S. and E. Van Cutsem (2005). Clinical signs, pathophysiology and management of skin toxicity during therapy with epidermal growth factor receptor inhibitors. *Ann Oncol* 16, 1425-33.
10. Playford, R.J., S. Ghosh, and A. Mahmood (2004). Growth factors and trefoil peptides in gastrointestinal health and disease. *Curr Opin Pharmacol* 4, 567-71.

11. Ozer, J., M. Ratner, M. Shaw, W. Bailey, and S. Schomaker (2008). The current state of serum biomarkers of hepatotoxicity. *Toxicology* 245, 194-205.
12. Wang, Y.M., S.C. Chai, C.T. Brewer, and T. Chen (2014). Pregnane X receptor and drug-induced liver injury. *Expert Opin Drug Metab Toxicol* 10, 1521-32.
13. Honig, P.K., R.L. Woosley, K. Zamani, D.P. Conner, and L.R. Cantilena Jr (1992). Changes in the pharmacokinetics and electrocardiographic pharmacodynamics of terfenadine with concomitant administration of erythromycin. *Clinical Pharmacology & Therapeutics* 52, 231-238.
14. Stephens, C., M.I. Lucena, and R.J. Andrade (2021). Genetic risk factors in the development of idiosyncratic drug-induced liver injury. *Expert Opinion on Drug Metabolism & Toxicology* 17, 153-169.
15. Zhou, Y., R. Tremmel, E. Schaeffeler, M. Schwab, and V.M. Lauschke (2022). Challenges and opportunities associated with rare-variant pharmacogenomics. *Trends in Pharmacological Sciences* 43, 852-865.
16. Uetrecht, J.P., *Adverse drug reactions*. Vol. 196. 2010: Springer.
17. Richard, A.M., R. Huang, S. Waidyanatha, P. Shinn, B.J. Collins, I. Thillainadarajah, C.M. Grulke, A.J. Williams, R.R. Lougee, R.S. Judson, K.A. Houck, M. Shobair, C. Yang, J.F. Rathman, A. Yasgar, S.C. Fitzpatrick, A. Simeonov, R.S. Thomas, K.M. Crofton, R.S. Paules, J.R. Bucher, C.P. Austin, R.J. Kavlock, and R.R. Tice (2021). The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem Res Toxicol* 34, 189-216.
18. Kleinstreuer, N.C., J. Yang, E.L. Berg, T.B. Knudsen, A.M. Richard, M.T. Martin, D.M. Reif, R.S. Judson, M. Polokoff, D.J. Dix, R.J. Kavlock, and K.A. Houck (2014). Phenotypic screening of the ToxCast chemical library to classify toxic and therapeutic mechanisms. *Nat Biotechnol* 32, 583-91.

19. Huang, R., M. Xia, S. Sakamuru, J. Zhao, S.A. Shahane, M. Attene-Ramos, T. Zhao, C.P. Austin, and A. Simeonov (2016). Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat Commun* 7, 10425.
20. Tasneem, A., L. Aberle, H. Ananth, S. Chakraborty, K. Chiswell, B.J. McCourt, and R. Pietrobon (2012). The database for aggregate analysis of ClinicalTrials.gov (AACT) and subsequent regrouping by clinical specialty. *PLoS One* 7, e33677.
21. Kuhn, M., I. Letunic, L.J. Jensen, and P. Bork (2016). The SIDER database of drugs and side effects. *Nucleic acids research* 44, D1075-D1079.
22. Banda, J.M., L. Evans, R.S. Vanguri, N.P. Tatonetti, P.B. Ryan, and N.H. Shah (2016). A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data* 3, 160026.
23. Tatonetti, N.P., P.Y. Patrick, R. Daneshjou, and R.B. Altman (2012). Data-driven prediction of drug effects and interactions. *Science translational medicine* 4, 125ra31-125ra31.
24. Chan, H.C.S., H. Shan, T. Dahoun, H. Vogel, and S. Yuan (2019). Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol Sci* 40, 592-604.
25. Hemmerich, J. and G.F. Ecker (2020). In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 10, e1475.
26. Inglese, J., R.L. Johnson, A. Simeonov, M. Xia, W. Zheng, C.P. Austin, and D.S. Auld (2007). High-throughput screening assays for the identification of chemical probes. *Nature chemical biology* 3, 466-479.
27. Szymański, P., M. Markowicz, and E. Mikiciuk-Olasik (2011). Adaptation of high-throughput screening in drug discovery—toxicological screening tests. *International journal of molecular sciences* 13, 427-452.
28. Kimlin, L., J. Kassis, and V. Virador (2013). 3D in vitro tissue models and their potential for drug screening. *Expert opinion on drug discovery* 8, 1455-1466.

29. Lo, Y.-C., S.E. Rensi, W. Torng, and R.B. Altman (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* 23, 1538-1546.
30. Van Norman, G.A. (2019). Limitations of animal studies for predicting toxicity in clinical trials: is it time to rethink our current approach? *JACC: Basic to Translational Science* 4, 845-854.
31. Raies, A.B. and V.B. Bajic (2016). In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdiscip Rev Comput Mol Sci* 6, 147-172.
32. Raies, A.B. and V.B. Bajic (2016). In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 6, 147-172.
33. Valerio Jr, L.G. (2009). In silico toxicology for the pharmaceutical sciences. *Toxicology and applied pharmacology* 241, 356-370.
34. Venkatapathy, R. and N.C.Y. Wang, *Developmental toxicity prediction*, in *Computational toxicology*. 2013, Springer. p. 305-340.
35. Roncaglioni, A., A.A. Toropov, A.P. Toropova, and E. Benfenati (2013). In silico methods to predict drug toxicity. *Current opinion in pharmacology* 13, 802-806.
36. Falk-Filipsson, A., A. Hanberg, K. Victorin, M. Warholm, and M. Wallén (2007). Assessment factors—applications in health risk assessment of chemicals. *Environmental research* 104, 108-127.
37. Martin, O.V., S. Martin, and A. Kortenkamp (2013). Dispelling urban myths about default uncertainty factors in chemical risk assessment—sufficient protection against mixture effects? *Environmental health* 12, 1-22.
38. Jack, J., J. Wambaugh, and I. Shah (2013). Systems toxicology from genes to organs. *Computational toxicology*, 375-397.
39. El-Masri, H., *Modeling for regulatory purposes (risk and safety assessment)*, in *Computational toxicology*. 2013, Springer. p. 297-303.

40. Sung, J.H., B. Srinivasan, M.B. Esch, W.T. McLamb, C. Bernabini, M.L. Shuler, and J.J. Hickman (2014). Using physiologically-based pharmacokinetic-guided “body-on-a-chip” systems to predict mammalian response to drug and chemical exposure. *Experimental biology and medicine* 239, 1225-1239.
41. Crump, K.S., C. Chen, W.A. Chiu, T.A. Louis, C.J. Portier, R.P. Subramaniam, and P.D. White (2010). What role for biologically based dose–response models in estimating Low-dose risk? *Environmental health perspectives* 118, 585-588.
42. Haber, L.T., A. Maier, Q. Zhao, J.S. Dollarhide, R.E. Savage, and M.L. Dourson (2001). Applications of mechanistic data in risk assessment: the past, present, and future. *Toxicological sciences* 61, 32-39.
43. Modi, S., M. Hughes, A. Garrow, and A. White (2012). The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug discovery today* 17, 135-142.
44. Jeliaskova, N., J. Jaworska, and A. Worth (2010). Open source tools for read-across and category formation. In *Silico Toxicology: Principles and Applications*, 408-445.
45. Dimitrov, S. and O. Mekenyan (2010). An introduction to read-across for the prediction of the effects of chemicals. In *silico toxicology: principles and applications*, 372-383.
46. Cherkasov, A., E.N. Muratov, D. Fourches, A. Varnek, Baskin, II, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha (2014). QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57, 4977-5010.
47. Sedykh, A., H. Zhu, H. Tang, L. Zhang, A. Richard, I. Rusyn, and A. Tropsha (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, 364-70.

48. Liu, J., K. Mansouri, R.S. Judson, M.T. Martin, H. Hong, M. Chen, X. Xu, R.S. Thomas, and I. Shah (2015). Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. *Chem Res Toxicol* 28, 738-51.
49. Ammad-ud-din, M., E. Georgii, M. Gonen, T. Laitinen, O. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski (2014). Integrative and personalized QSAR analysis in cancer by kernelized Bayesian matrix factorization. *J Chem Inf Model* 54, 2347-59.
50. Yamane, J., S. Aburatani, S. Imanishi, H. Akanuma, R. Nagano, T. Kato, H. Sone, S. Ohsako, and W. Fujibuchi (2016). Prediction of developmental chemical toxicity based on gene networks of human embryonic stem cells. *Nucleic Acids Res* 44, 5515-28.
51. Capuzzi, S.J., R. Politi, O. Isayev, S. Farag, and A. Tropsha (2016). QSAR Modeling of Tox21 Challenge Stress Response and Nuclear Receptor Signaling Toxicity Assays. *Frontiers in Environmental Science* 4.
52. Zhang, J., D. Mucs, U. Norinder, and F. Svensson (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. *J Chem Inf Model* 59, 4150-4158.
53. Mayr, A., G. Klambauer, T. Unterthiner, and S. Hochreiter (2016). DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* 3.
54. Idakwo, G., S. Thangapandian, J.t. Luttrell, Z. Zhou, C. Zhang, and P. Gong (2019). Deep Learning-Based Structure-Activity Relationship Modeling for Multi-Category Toxicity Classification: A Case Study of 10K Tox21 Chemicals With High-Throughput Cell-Based Androgen Receptor Bioassay Data. *Front Physiol* 10, 1044.
55. Matsuzaka, Y. and Y. Uesawa (2020). Molecular Image-Based Prediction Models of Nuclear Receptor Agonists and Antagonists Using the DeepSnap-Deep Learning Approach with the Tox21 10K Library. *Molecules* 25.
56. Wu, L., R. Huang, I.V. Tetko, Z. Xia, J. Xu, and W. Tong (2021). Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chem Res Toxicol* 34, 541-549.

57. Wang, Z., N.R. Clark, and A. Ma'ayan (2016). Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 32, 2338-45.
58. Dudley, J.T., T. Deshpande, and A.J. Butte (2011). Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 12, 303-11.
59. Polishchuk, P. (2017). Interpretation of Quantitative Structure-Activity Relationship Models: Past, Present, and Future. *J Chem Inf Model* 57, 2618-2639.
60. Xue, C.X., R.S. Zhang, H.X. Liu, X.J. Yao, M.C. Liu, Z.D. Hu, and B.T. Fan (2004). QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *J Chem Inf Comput Sci* 44, 1693-700.
61. Naboulsi, I., A. Aboulmouhajir, L. Kouisni, F. Bekkaoui, and A. Yasri (2018). Combining a QSAR Approach and Structural Analysis to Derive an SAR Map of Lyn Kinase Inhibition. *Molecules* 23.
62. Ahamed, T.S., V.K. Rajan, and K. Muraleedharan (2019). QSAR modeling of benzoquinone derivatives as 5-lipoxygenase inhibitors. *Food Science and Human Wellness* 8, 53-62.
63. Heo, S., U. Safder, and C. Yoo (2019). Deep learning driven QSAR model for environmental toxicology: Effects of endocrine disrupting chemicals on human health. *Environ Pollut* 253, 29-38.
64. Lorberbaum, T., M. Nasir, M.J. Keiser, S. Vilar, G. Hripcsak, and N.P. Tatonetti (2015). Systems pharmacology augments drug safety surveillance. *Clin Pharmacol Ther* 97, 151-8.
65. Hao, Y., K. Quinnes, R. Realubit, C. Karan, and N.P. Tatonetti (2018). Tissue-Specific Analysis of Pharmacological Pathways. *CPT Pharmacometrics Syst Pharmacol* 7, 453-463.
66. Urbanowicz, R.J., M. Meeker, W. La Cava, R.S. Olson, and J.H. Moore (2018). Relief-based feature selection: Introduction and review. *J Biomed Inform* 85, 189-203.

67. Urbanowicz, R.J., R.S. Olson, P. Schmitt, M. Meeker, and J.H. Moore (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. *J Biomed Inform* 85, 168-188.
68. Gilson, M.K., T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong (2016). BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* 44, D1045-D1053.
69. Wishart, D.S., Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46, D1074-D1082.
70. Willighagen, E.L., J.W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Cherto, O. Spjuth, G. Torrance, C.T. Evelo, R. Guha, and C. Steinbeck (2017). The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 9, 33.
71. Kumar, R.D., L.W. Chang, M.J. Ellis, and R. Bose (2013). Prioritizing Potentially Druggable Mutations with dGene: An Annotation Tool for Cancer Genome Sequencing Data. *PLoS One* 8, e67980.
72. Armstrong, J.F., E. Faccenda, S.D. Harding, A.J. Pawson, C. Southan, J.L. Sharman, B. Campo, D.R. Cavanagh, S.P.H. Alexander, A.P. Davenport, M. Spedding, J.A. Davies, and I. Nc (2020). The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res* 48, D1006-D1021.
73. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, and J.T. Eppig (2000). Gene Ontology: tool for the unification of biology. *Nature genetics* 25, 25-29.

74. Gene Ontology, C. (2021). The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res* 49, D325-D334.
75. Jassal, B., L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorsler, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio (2020). The reactome pathway knowledgebase. *Nucleic Acids Res* 48, D498-D503.
76. Pinero, J., J.M. Ramirez-Angueta, J. Sauch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L.I. Furlong (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 48, D845-D855.
77. Suzuki, R. and H. Shimodaira (2006). Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540-2.
78. Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318-1330.
79. Davis, A.P., C.J. Grondin, R.J. Johnson, D. Sciaky, J. Wieggers, T.C. Wieggers, and C.J. Mattingly (2020). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res*.
80. Alonso, A., X. Yin, N.S. Roetker, J.W. Magnani, R.A. Kronmal, P.T. Ellinor, L.Y. Chen, S.A. Lubitz, R.L. McClelland, D.D. McManus, E.Z. Soliman, R.R. Huxley, S. Nazarian, M. Szklo, S.R. Heckbert, and E.J. Benjamin (2014). Blood lipids and the incidence of atrial fibrillation: the Multi-Ethnic Study of Atherosclerosis and the Framingham Heart Study. *J Am Heart Assoc* 3, e001211.
81. Croquelois, A., A.A. Domenighetti, M. Nemir, M. Lepore, N. Rosenblatt-Velin, F. Radtke, and T. Pedrazzini (2008). Control of the adaptive response of the heart to stress via the Notch1 receptor pathway. *J Exp Med* 205, 3173-85.
82. Zuercher, M., K.B. Kern, J.H. Indik, M. Loedl, R.W. Hilwig, W. Ummenhofer, R.A. Berg, and G.A. Ewy (2011). Epinephrine improves 24-hour survival in a swine model of

- prolonged ventricular fibrillation demonstrating that early intraosseous is superior to delayed intravenous administration. *Anesth Analg* 112, 884-90.
83. Shaked, I., M.A. Oberhardt, N. Atias, R. Sharan, and E. Ruppin (2016). Metabolic Network Prediction of Drug Side Effects. *Cell Syst* 2, 209-13.
84. Society for Advancement of AOPs. 2021 4/24/2021 [cited 2021 Sep 17, 2021]; Release 2.4:[Available from: <http://aopwiki.org>].
85. Vrijenhoek, N. *IKK complex inhibition leading to liver injury*. 2021 Jun 04, 2021 [cited 2021 Sep 17, 2021]; Available from: <https://aopwiki.org/aops/278>.
86. Yangh, H., K.E. Snijders, and M. Niemeijer. *Inhibition of N-linked glycosylation leads to liver injury*. 2021 Jun 04, 2021 [cited 2021 Sep 17, 2021]; Available from: <https://aopwiki.org/aops/285>.
87. Choi, J., N. Chatterjee, J. Jeong, J.-y. Rho, E.-Y. Kim, S.M. Oh, N.I. Garcia-Reyero, E.J. Perkins, and L.D. Burgoon. *Peroxisome proliferator-activated receptors γ inactivation leading to lung fibrosis*. 2021 Sep 16, 2021 [cited 2021 Sep 17, 2021]; Available from: <https://aopwiki.org/aops/206>.
88. Zhang, W., R. An, Q. Li, L. Sun, X. Lai, R. Chen, D. Li, and S. Sun (2020). Theaflavin TF3 Relieves Hepatocyte Lipid Deposition through Activating an AMPK Signaling Pathway by targeting Plasma Kallikrein. *J Agric Food Chem* 68, 2673-2683.
89. Zou, X., P. Ramachandran, T.J. Kendall, A. Pellicoro, E. Dora, R.L. Aucott, K. Manwani, T.Y. Man, K.E. Chapman, N.C. Henderson, S.J. Forbes, S.P. Webster, J.P. Iredale, B.R. Walker, and Z. Michailidou (2018). 11Beta-hydroxysteroid dehydrogenase-1 deficiency or inhibition enhances hepatic myofibroblast activation in murine liver fibrosis. *Hepatology* 67, 2167-2181.
90. Carta, F. and C.T. Supuran (2013). Diuretics with carbonic anhydrase inhibitory action: a patent and literature review (2005 - 2013). *Expert Opin Ther Pat* 23, 681-91.

91. Chandrasekharan, B., V. Bala, V.L. Kolachala, M. Vijay-Kumar, D. Jones, A.T. Gewirtz, S.V. Sitaraman, and S. Srinivasan (2008). Targeted deletion of neuropeptide Y (NPY) modulates experimental colitis. *PLoS One* 3, e3304.
92. Oliveira, T.L., N. Candeia-Medeiros, P.M. Cavalcante-Araujo, I.S. Melo, E. Favaro-Pipi, L.A. Fatima, A.A. Rocha, L.R. Goulart, U.F. Machado, R.R. Campos, and R. Sabino-Silva (2016). SGLT1 activity in lung alveolar cells of diabetic rats modulates airway surface liquid glucose concentration and bacterial proliferation. *Sci Rep* 6, 21752.
93. Ribeiro, M.T., S. Singh, and C. Guestrin, "Why Should I Trust You?", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. 2016. p. 1135-1144.
94. Shrikumar, A., P. Greenside, and A. Kundaje. *Learning important features through propagating activation differences*. in *International conference on machine learning*. 2017. PMLR.
95. Du, M., N. Liu, and X. Hu (2019). Techniques for interpretable machine learning. *Communications of the ACM* 63, 68-77.
96. Ma, J., M.K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat Methods* 15, 290-298.
97. Kuenzi, B.M., J. Park, S.H. Fong, K.S. Sanchez, J. Lee, J.F. Kreisberg, J. Ma, and T. Ideker (2020). Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* 38, 672-684 e6.
98. Lin, C.H. and O. Lichtarge (2021). Using Interpretable Deep Learning to Model Cancer Dependencies. *Bioinformatics*.
99. Elmarakeby, H.A., J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S.H. AlDubayan, K. Salari, S. Kregel, C. Richter, T.E. Arnoff, J. Park, W.C. Hahn, and E.M. Van Allen (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature* 598, 348-352.

100. Subramanian, A., R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, J. Gould, J.F. Davis, A.A. Tubelli, J.K. Asiedu, D.L. Lahr, J.E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I.C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O.M. Enache, F. Piccioni, S.A. Johnson, N.J. Lyons, A.H. Berger, A.F. Shamji, A.N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D.Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N.S. Gray, P.A. Clemons, S. Silver, X. Wu, W.N. Zhao, W. Read-Button, X. Wu, S.J. Haggarty, L.V. Ronco, J.S. Boehm, S.L. Schreiber, J.G. Doench, J.A. Bittker, D.E. Root, B. Wong, and T.R. Golub (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171, 1437-1452 e17.
101. Cereto-Massague, A., M.J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallve, and G. Pujadas (2015). Molecular fingerprint similarity search in virtual screening. *Methods* 71, 58-63.
102. Hao, Y. and J.H. Moore (2021). TargetTox: A Feature Selection Pipeline for Identifying Predictive Targets Associated with Drug Toxicity. *J Chem Inf Model* 61, 5386-5394.
103. Bach, S., A. Binder, G. Montavon, F. Klauschen, K.R. Muller, and W. Samek (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* 10, e0130140.
104. Montavon, G., A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller (2019). Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193-209.
105. Romano, J.D., Y. Hao, and J.H. Moore (2022). Improving QSAR Modeling for Predictive Toxicology using Publicly Aggregated Semantic Graph Data and Graph Neural Networks. *Pac Symp Biocomput* 27, 187-198.
106. Sarwar, M., S. Sandberg, P.-A. Abrahamsson, and J.L. Persson. *Protein kinase A (PKA) pathway is functionally linked to androgen receptor (AR) in the progression of prostate cancer*. in *Urologic Oncology: Seminars and Original Investigations*. 2014. Elsevier.

107. Lamb, L.E., J.C. Zarif, and C.K. Miranti (2011). The androgen receptor induces integrin $\alpha 6\beta 1$ to promote prostate tumor cell survival via NF- κ B and Bcl-xL Independently of PI3K signaling. *Cancer research* 71, 2739-2749.
108. Wang, X., M.M. Docanto, H. Sasano, C. Kathleen Cuningham Foundation Consortium for Research into Familial Breast, C. Lo, E.R. Simpson, and K.A. Brown (2015). Prostaglandin E2 inhibits p53 in human breast adipose stromal cells: a novel mechanism for the regulation of aromatase in obesity and breast cancer. *Cancer Res* 75, 645-55.
109. Iorga, A. and L. Dara (2019). Cell death in drug-induced liver injury. *Adv Pharmacol* 85, 31-74.
110. Chen, Y., Y. Wang, Y. Zhuang, F. Zhou, and L. Huang (2012). Mifepristone increases the cytotoxicity of uterine natural killer cells by acting as a glucocorticoid antagonist via ERK activation. *PLoS One* 7, e36413.
111. Yao, X.P., T.Y. Jiao, Y.M. Jiang, S.C. Fan, Y.Y. Zhao, X. Yang, Y. Gao, F. Li, Y.Y. Zhou, P.P. Chen, M. Huang, and H.C. Bi (2022). PXR mediates mifepristone-induced hepatomegaly in mice. *Acta Pharmacol Sin* 43, 146-156.
112. Srivastava, M.D., A. Thomas, B.I. Srivastava, and J.H. Check (2007). Expression and modulation of progesterone induced blocking factor (PIBF) and innate immune factors in human leukemia cell lines by progesterone and mifepristone. *Leuk Lymphoma* 48, 1610-7.
113. Brentnall, M., L. Rodriguez-Menocal, R.L. De Guevara, E. Cepero, and L.H. Boise (2013). Caspase-9, caspase-3 and caspase-7 have distinct roles during intrinsic apoptosis. *BMC Cell Biol* 14, 32.
114. Yosefzon, Y., D. Soteriou, A. Feldman, L. Kostic, E. Koren, S. Brown, R. Ankawa, E. Sedov, F. Glaser, and Y. Fuchs (2018). Caspase-3 Regulates YAP-Dependent Cell Proliferation and Organ Size. *Mol Cell* 70, 573-587 e4.
115. Wang, C. and R.J. Youle (2009). The role of mitochondria in apoptosis*. *Annu Rev Genet* 43, 95-118.

116. Albenzi, B.C. (2019). What Is Nuclear Factor Kappa B (NF-kappaB) Doing in and to the Mitochondrion? *Front Cell Dev Biol* 7, 154.
117. Bonifaz, L., M. Cervantes-Silva, E. Ontiveros-Dotor, E. Lopez-Villegas, and F. Sanchez-Garcia (2014). A Role For Mitochondria In Antigen Processing And Presentation. *Immunology*.
118. Yin, S. and B. Gao (2010). Toll-like receptor 3 in liver diseases. *Gastroenterol Res Pract* 2010.
119. Guo, J. and S.L. Friedman (2010). Toll-like receptor 4 signaling in liver injury and hepatic fibrogenesis. *Fibrogenesis Tissue Repair* 3, 21.
120. Grulke, C.M., A.J. Williams, I. Thillanadarajah, and A.M. Richard (2019). EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Comput Toxicol* 12.
121. Hao, Y., J.D. Romano, and J.H. Moore (2022). Knowledge-guided deep learning models of drug toxicity improve interpretation. *Patterns* 3, 100565.
122. Romano, J.D., Y. Hao, J.H. Moore, and T.M. Penning (2022). Automating Predictive Toxicology Using ComptoxAI. *Chemical Research in Toxicology*.
123. Romano, J.D., Y. Hao, and J.H. Moore. *Improving QSAR Modeling for Predictive Toxicology using Publicly Aggregated Semantic Graph Data and Graph Neural Networks*. in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022*. 2021. World Scientific.
124. Kingma, D.P. and J. Ba (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
125. Quinn, N.P., L. García-Gutiérrez, C. Doherty, A. von Kriegsheim, E. Fallahi, D.B. Sacks, and D. Matallanas (2021). IQGAP1 is a scaffold of the core proteins of the hippo pathway and negatively regulates the pro-apoptotic signal mediated by this pathway. *Cells* 10, 478.

126. Eadie, L., T. Hughes, and D. White (2014). Interaction of the efflux transporters ABCB1 and ABCG2 with imatinib, nilotinib, and dasatinib. *Clinical Pharmacology & Therapeutics* 95, 294-306.
127. Veith, A. and B. Moorthy (2018). Role of cytochrome P450s in the generation and metabolism of reactive oxygen species. *Current opinion in toxicology* 7, 44-51.
128. Méndez-Lucio, O., M. Ahmad, E.A. del Rio-Chanona, and J.K. Wegner (2021). A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence* 3, 1033-1039.
129. Zeng, Y., X. Chen, Y. Luo, X. Li, and D. Peng (2021). Deep drug-target binding affinity prediction with multiple attention blocks. *Briefings in bioinformatics* 22, bbab117.
130. Monteiro, N.R., J.L. Oliveira, and J.P. Arrais (2022). DTITR: End-to-end drug–target binding affinity prediction with transformers. *Computers in Biology and Medicine* 147, 105772.
131. Sedykh, A., H. Zhu, H. Tang, L. Zhang, A. Richard, I. Rusyn, and A. Tropsha (2011). Use of in vitro HTS-derived concentration–response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environmental health perspectives* 119, 364-370.
132. Varshneya, M., X. Mei, and E.A. Sobie (2021). Prediction of arrhythmia susceptibility through mathematical modeling and machine learning. *Proceedings of the National Academy of Sciences* 118, e2104019118.
133. Guengerich, F.P. (2011). Mechanisms of drug toxicity and relevance to pharmaceutical development. *Drug metabolism and pharmacokinetics* 26, 3-14.
134. Riley, A.L. and S. Kohut, *Drug Toxicity*, in *Encyclopedia of Psychopharmacology*, I.P. Stolerman, Editor. 2010, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 441-441.
135. Greene, C.S., A. Krishnan, A.K. Wong, E. Ricciotti, R.A. Zelaya, D.S. Himmelstein, R. Zhang, B.M. Hartmann, E. Zaslavsky, and S.C. Sealfon (2015). Understanding

multicellular function and disease with human tissue-specific networks. *Nature genetics* 47, 569-576.

136. Wong, A.K., A. Krishnan, and O.G. Troyanskaya (2018). GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. *Nucleic acids research* 46, W65-W70.
137. Lee, Y.-s., A.K. Wong, A. Tadych, B.M. Hartmann, C.Y. Park, V.A. DeJesus, I. Ramos, E. Zaslavsky, S.C. Sealfon, and O.G. Troyanskaya (2018). Interpretation of an individual functional genomics experiment guided by massive public data. *Nature methods* 15, 1049-1052.