Computational approaches for Drug-induced liver injury (DILI) prediction: state of the art and challenges

Olivier J. M. Béquignon[1], Gopal Pawar[2,3], Bob van de Water[1], Mark T. D. Cronin[2], Gerard J. P. van Westen[1]

[1] Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands

[2] School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, United Kingdom

[3] Currently at the Institute of Clinical Sciences, Pharmacy, University of Birmingham, Birmingham, United Kingdom

Author email addresses: OJM Bequignon - o.j.m.bequignon@lacdr.leidenuniv.nl , G Pawar - g.pawar@ljmu.ac.uk, B vd Water - water_b@lacdr.leidenuniv.nl, MTD Cronin - m.t.cronin@ljmu.ac.uk, GJP v Westen – gerard@lacdr.leidenuniv.nl

KEYWORDS: machine learning, QSAR, adverse event, toxicity, drug safety, in silico, modelling

ABSTRACT

Drug-induced liver injury (DILI) is one of the prevailing causes of fulminant hepatic failure. It is estimated that three idiosyncratic drug reactions out of four result in liver transplantation or death. Additionally, DILI is the most common reason for withdrawal of an approved drug from the market. Therefore, the development of methods for the early identification of hepatotoxic drug candidates is of crucial importance. This review focuses on the current state of cheminformatics strategies being applied for the early *in silico* prediction of DILI. Herein, we discuss key issues associated with DILI modelling in terms of the data size, imbalance and quality, complexity of mechanisms, and the different levels of hepatotoxicity to model going from general hepatotoxicity to the molecular initiating events of DILI.

INTRODUCTION

Drug-induced liver injury (DILI) refers to hepatotoxicity resulting from adverse reactions caused by drugs or their reactive metabolites and toxic chemical entities. DILI is a major concern as it is one of the leading causes of acute liver failure in the world, accounting for more than 50% of cases in the US[1]. Additionally, a recent study showed that DILI is responsible for more than 20% of the withdrawals of approved drugs from the market due to toxicity [2–4]. This is an on-going problem, there have been at least eight withdrawals of drugs due to DILI from 1997 to 2016 alone: tolcapone, troglitazone, trovafloxacin, bromfenac, nefazodone, ximelagatran, lumiracoxib and sitaxentan[5]. Moreover, hepatotoxicity is also a major reason for the failure of candidates in the drug discovery process[6]. These reasons underscore the need for the accurate prediction of the risk of DILI for bioactive compounds. DILI itself is complex, it comprises a broad set of effects which can be further characterised in several ways, either by the type of hepatotoxicity (physiological effect) or by whether the effect is dose-dependent or not.

With regard to hepatotoxicity, three types or patterns may be observed. Firstly, hepatocellular injury which is the result of biochemical perturbations of the cell culminating in severe cellular malfunction or cell death, the latter resulting in formation of scaring tissue. It comprises steatosis, necrosis and cirrhosis and is characterised by the release of hepatocellular enzymes (e.g. alanine transferase (ALT) and aspartate transaminase (AST)). Secondly, cholestatic injury is the result of an impairment of the biliary system caused either by bile stasis (i.e. the accumulation of bile in the bile ducts), portal inflammation or proliferation or injury of bile ducts. It is usually characterised by elevated levels of alkaline phosphatase (ALP) and γ-glutamyl transpeptidase (GGT). Finally, mixed hepatocellular-cholestatic injury, which occurs rarely in other forms of acute liver disease, usually shows prominent hepatocyte necrosis and inflammation as well as marked bile stasis. It is characterised by the elevation of both ALT and ALP.

DILI itself may also be categorised into two subtypes. The first type, called intrinsic DILI (itDILI), is dose-dependent and is modulated by the presence of key compound substructures and its effects are reversed after discontinuation of drug administration. These reasons make it quite predictable[7]. The second type is idiosyncratic DILI (iDILI), which is very rare as it only occurs in 1:1,000 to 1:100,000 patients exposed to the drug[8]. iDILI is associated with poor prognosis and does not show any dose-response relationship. Because it is host-dependent[9,10], iDILI can be the result of either immunological effects (i.e. allergic reactions) or metabolic effects which makes it more unpredictable[11] and a considerable challenge for drug development and safety.

These problems emphasise the importance of the early detection of hepatotoxic compounds in the drug discovery process in order to reduce attrition rates and to increase drug safety. However, a major obstacle to the development of comprehensive tools for the early detection of iDILI is primarily the lacking predictivity of the existing animal studies and secondly its complexity, ranging from the variety of its effects but also from the diversity of factors affecting susceptibility to iDILI. Additionally, drug metabolism and pharmacokinetics (DMPK) aspects, including local and intracellular concentration, are difficult to evaluate and predict. Effects of iDILI include elevations in serum transaminases, jaundice, acute liver failure or chronic liver dysfunction. Factors affecting iDILI include age, gender, ethnicity, genetic polymorphism, use of other medication or pre-existing liver disease[12,13]. Additionally, the development and mechanisms of iDILI are poorly understood making its early detection, and therefore its prediction, a challenge[14,15]. A detailed summary of these mechanisms lies outside the scope of this review and the reader is referred to the works of Fraser *et al.*[16] and of Noureddin and Kaplowitz[17] for comprehensive information on DILI mechanisms. Nonetheless, a wide range of predictive models have been established for the prediction of DILI and can be divided among quantitative adverse outcome pathways (qAOPs)[18], metabolomics[19], cheminformatics [14,20], pharmacokinetic-pharmacodynamics (PK-PD) modelling[21], dynamical pathway modelling with ordinary differential equation (ODE) models[22] and

multi-scale approaches modelling DILI with systems biology approaches[23].

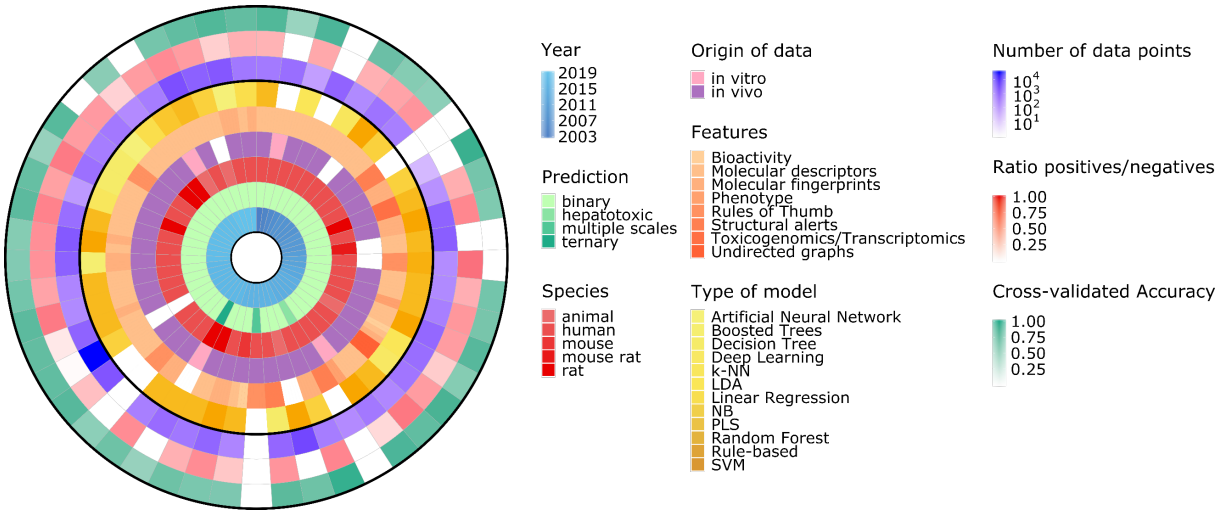

Figure 1: Visual summary of in silico models for liver toxicity prediction

The focus of this work is to characterise the application and scope of published cheminformatics models for DILI and to highlight their relevance, with a particular focus on machine learning.

APPROACHES TO PREDICTING DILI RISK

Better understanding of the underlying mechanisms of DILI, as well as better annotation of the risk associated with drug structures is key for the development of more accurate and valuable predictive models[20]. Additionally there is no evidence that the mechanisms through which iDILI occurs are different than itDILI[24,25]. Thus, the focus of DILI research has been to identify reported clinical cases of hepatotoxicity. For instance, such information was compiled by Ludwig and Axelsen[26], who created a list of 150 compounds associated with their adverse events. This compilation did not account for the difference(s) between itDILI and iDILI but was one of the first exhaustive lists of hepatotoxic drugs to link phenotypic outcomes in human.

A more recent study classified a list of 611 compounds using high content image screening (HCS) on human cells and compared the findings to conventional assays[27]. The compounds were classified as either "severely", "moderately" or non-toxic and laid the foundation for the use of *in vitro* data as a surrogate for the prediction of clinical outcomes. Other sources of hepatotoxicity-related compounds come from medicines regulatory agencies and post-marketing data. For instance, Suzuki *et al.* compiled adjudicated cases of DILI reported from the literature resulting from drugs that had been suspended or withdrawn from the market[28] and Chen *et al.* annotated compounds based on information provided by the United States Food and Drug Administration (FDA)[4]. The first version of the latter organised compounds into three categories: no-DILI Concern compounds, for which no hepatotoxicity had been

observed, Less-DILI Concern, which caused only mild hepatotoxicity (i.e. steatosis, cholestasis and increase in liver aminotransferases) and Most-DILI concern, which were associated with severe hepatotoxicity[4]. In a later revision, called DILIRank[29], the data were curated based on causality evidence. This allowed for the separation of compounds for which association with hepatotoxicity was not supported by sufficient data and allowed for the creation of a new class of compounds (i.e. Ambiguous DILI Concern) consisting of the compounds of the Most and Less DILI Concern classes of the previous version of DILIRank for which no strong evidence of causality was observed.

Fourches *et al.* used text-mining approaches on the titles and abstracts of a collection of articles to identify 902 compounds associated with drug-induced liver effects[30]. Based on these different approaches to annotate compounds, Kotsampasakou *et al.* aggregated the data from 9 datasets and applied extensive curation techniques[31]. Multiple datasets have been published[32] either derived from clinical and/or post-marketing sources, from *in vitro/in vivo* experiments or aggregated from different types of sources (Table 1). However, the published data suffer from two major limitations: data size and imbalance in both the positive and negative DILI group compounds which would bias the outcome of the analysis.

LIMITED DATASET SIZES HAMPERS PROPER MODEL VALIDATION

As a consequence of the nature of the datasets described above, the majority of existing published models for DILI are binary classification models (Table 2). Of these, only one, by Cheng and Dixon, focused exclusively on the prediction of reported itDILI in humans using a set of 382 compounds related to 25 2D molecular descriptors selected with a Monte-Carlo regression algorithm[7]. The leave-10%-out cross-validated random forest model developed had very high specificity and reasonable sensitivity (0.90 and 0.78 respectively). Although similar performance was observed with the test set, its size was quite limited as it only included 23 positive compounds and 31 negatives. Similarly Cruz-Monteagudo *et al.* developed general hepatotoxicity binary classification models from a set of 74 compounds using Radial Distribution Function (RDF) descriptors[33]. Even though the performance of the best performing model was consistent between the cross-validation and external validation sets (0.86 and 0.82 respectively), the validation set was small comprising only 13 hepatotoxic compounds and no negatives.

Table 1: Published classifications of drugs for DILI risk

Year	*Reference*		*origin of data*	*number of compounds*	*endpoint*
1983	26	Ludwig and Axelsen	Compilation of published data	150	Morphological endpoints
1999	34	Zimmerman	Compilation of published data	~500 hepatotoxic drugs	
2005	35	Guo *et al.*	Compilation of public data	175 drugs	0: no information about hepatotoxicity 1: no significant liver damage reported 2: multiple cases reports or significant injury 3: clear literature evidence of life-threatening hepatotoxicity
2006	27	O'Brien *et al.*	*in vitro* cell-based data	381 (42 +/102 ~/237 –)	Severely, moderately and non-toxic
2007	36–38	ToxCast	*in vitro* data	3799	Biochemical properties based on HTS assays, cell-based phenotypic assays, and genomic and

Year	*Reference*		*origin of data*	*number of compounds*	*endpoint*
					metabolomic analyses of cells
2008	39	Xu *et al.*	Drug labels, expert opinion	344 (200 +/144 –)	Hepatotoxic and non-hepatotoxic
2010	40	Ekins *et al.*	Clinical data for hepatotoxicity	532 (272 +/ 260 –)	Based on Xu *et al.* [39]
2010	30	Fourches *et al.*	Text mining	951 compounds	Liver effects in humans, rodents or non-rodents
2010	41	Greene *et al.*	Compilation of published data	1,266	Human or animal-only hepatotoxicity, weak or no evidence
2010	42	Rodgers *et al.*	FDA reports database	395 (76 +/ 319 –)	Compounds (not) associated with ALT or AST elevation or with combined score
2010	43,44	SIDER database	Compilation of public data	1,430 drugs	Association with 5,868 hepatotoxic side effects
2010	28	Suzuki *et al.*	Compilation of data from regulatory agencies	473 hepatotoxic drugs	Drugs causing overall injury, acute liver failure and suspended/withdrawn
2011	4	Chen *et al.*	FDA-approved labels	287 drugs	Most-DILI Concern Less-DILI Concern No DILI Concern
2011	45	Liew *et al.*	Micromedex reports of adverse reactions	1,274 compounds	1: transient and asymptomatic liver function abnormalities 2: liver function abnormalities and hyperbilirubinemia 3: hepatitis, jaundice and cholestasis 4: fulminant hepatitis and liver failure 5: fatality
2011	46	Liu *et al.*	SIDER database	888 drugs	Association with 13 hepatotoxic side effects
2011	47	Low *et al.*	*in vivo* toxicogenomics on rats	127 (53 +/74 –)	Hepatotoxic and non-hepatotoxic
2012	48	Sakatis *et al.*	Physician's Desk reference	223 (113 +/ 110 –)	Hepatotoxic and non-hepatotoxic
2013	49	Liver Toxicity Knowledge Base	FDA-approved labels	195 (113 +/ 82 –)	Most-DILI Concern No DILI Concern
2013	50	LiverTox	Compilation of published data	~1,200 hepatotoxic drugs, dietary supplements and herbal products	
2014	51	Zhu & Kruhlak	Post-marketing safety data	2,029 (662 +/ 1367 –)	Hepatotoxic and non-hepatotoxic
2016	52	DILIrank	FDA-approved labels	1036 drugs	Verified Most-DILI Concern Verified Less-DILI Concern Verified No DILI Concern Ambiguous DILI Concern
2016	53	Mulliner *et al.*	Compilation of public data	921 (519 +/ 402 –)	Hierarchical classification in 21 endpoints
2016	54,55	Tox21	In vitro data	~ 10,000	Biochemical properties based on HTS assays
2016		eTOX	*in vitro* and *in vivo* data	1947	In-life observations, gross necropsies, histopathology and laboratory values (e.g. clinical chemistry, haematology and urinalysis)
2017	31	Kotsampasakou *et al.*	Compilation of published data	966 (500 +/ 466 –)	Hepatotoxic and non-hepatotoxic
2018	56	Ai *et al.*	Zhu & Kruhlak [51], FDA Orange Book	1,241 (683 +/ 558 –)	Hepatotoxic and non-hepatotoxic

Although the metrics indicate that Cheng and Dixon's and Cruz-Monteagudo *et al.*'s models performed well, one has to consider that a phenotypic readout such as general hepatotoxicity is the integrated result of many signalling pathways (e.g. oxidative stress and NRF2 pathway[57], unfolded protein response, DNA damage response and mitochondrial toxicity[17]). For each pathway, protein-protein interactions, as well as gene expression or gene and protein degradation could be disturbed, adding up to a multitude of different modes of actions by which a compound could induce toxicity. Thus, building general hepatotoxicity models from a rather small number of diverse compounds increases the difficulty to make reliable generalisations based on compound structures when considering all the possible toxicity modes of action that could be triggered. Xu *et al*. exemplified such a phenomenon and showed that an increase

of the size of the training set improved not only the accuracy of models but also reduced their variability[58]. Additionally, the limited size of external test sets (Table 1) makes the interpretation of the validation of hepatotoxicity prediction models difficult since only a small fraction of hepatotoxicity mechanisms may be validated. The ideal validation set should comprise at least as many compounds as there are ways to disturb the processes involved in these pathways. However, the aggregation of such a dataset is, at this time, not possible. Nevertheless, sizes of both training sets and evaluation sets have been increasing (Table 2), notably through the aggregation and careful data curation of multiple datasets[59] but also through the United States Environmental Protection Agency's (EPA) ToxCast[36–38] and the multi-agency Tox21[54,55] open-data initiatives and the European eTOX[60–63] and eTRANSAFE consortia. These consortia have gathered pharmaceuticals, data curators, modelers and software developers aiming at building a shared and mineable database of preclinical (eTOX) and clinical (eTRANSAFE) toxicity data to enable more effective read-across and predictive modelling of safety endpoints.

DATASET IMBALANCE

The second limitation of published datasets is the imbalance of the validation sets (e.g. in [33,64–68] in Table 2). These datasets, where either only hepatotoxic compounds are represented or fewer than 10% of compounds are non-hepatotoxicants, do not allow for a proper estimation of the specificity of the models. From the perspective of the training set, the imbalance of the data has been a major challenge to overcome in the prediction of hepatotoxicity: we identified eight articles in which the ratio of non-hepatotoxic compounds considered represented less than 40% of the training set[7,13,42,69–73]. The opposite trend was observed in six articles where hepatotoxic compounds represented less than 40% of the training set[68,74–78]. Although building a robust model on an imbalanced dataset is possible, the performance decreases significantly when the number of individuals in the minority class approaches, or becomes, less than 10%. Whilst imbalanced sets affect the robustness of a model, they may better represent the distribution of compounds or dugs observed in real life. This is relevant for the work of Lu *et al.*, who predicted the general hepatotoxicity of compounds based on the profiles of their predicted metabolites[69], where 64 hepatotoxic and 3,339 non-hepatotoxic compounds were considered – the minority class representing about 2% of the entire dataset. The strategies generally adopted to counteract the systematic prediction of compounds to belong to the majority class are (i) undersampling of the majority class, (ii) oversampling of the minority class[79], (iii) bagging, (iv) boosting, (v) cost-sensitive learning and (vi) hybrid methods[80,81]. In their work, Lu *et al.* used the Synthetic Minority Oversampling Technique (SMOTE) algorithm[79] to correct for this data imbalance yielding a cross-validated balanced accuracy of 0.60 when predicting hepatotoxicity from predicted metabolites[69]. The application of such meta-classifiers in the prediction of hepatotoxicity is quite recent since only five other works have used them since 2015[13,72,77,82–84]. It is worth noting that a comparison of the behaviour of meta-classifiers has been performed on few selected imbalanced drug-induced cholestasis datasets[85]. Bagging[86] has the worst

performance as it does not balance or weight the two classes, threshold selection performed better than bagging but gave lower sensitivity than when using stratified bagging, cost sensitive classifier or Meta-Cost[87]. The authors emphasised the versatility of the stratified bagging technique despite its computational cost when extensive resampling has to be performed.

Early DILI Prediction Strategies

Among the different *in silico* models that have been developed for the prediction of hepatotoxicity, four main groups of models can be identified based on the features, properties or data the prediction models are built upon: (i) structural alerts, (ii) rules of thumb, (iii) molecular descriptors and (iv) *in vitro* data. These are described in detail below.

Structural Alerts

Structural alerts are specific substructures of molecules generally associated with hepatotoxicity. Structural alerts are generally developed by experts in toxicology who consider not only toxicological data but also the underlying mechanisms of toxicity, as well as chemical reactivity and biotransformation through metabolism.

One of the first approaches to determining such alerts for DILI utilised a four-stage process[41]. A dataset of 1,266 compounds associated with *in vivo* human DILI was aggregated from the literature. Candidate structural classes were derived from these compounds by experts through well-characterised and previously published relationships between compound structures and hepatoxicity. Then these classes were refined by the development of structure-activity relationships (SAR) for which sufficient evidence was available. Finally, the 38 structural alerts classes identified, such as tetracyclines and thiophenes, were validated against an in-house dataset from Pfizer consisting of 626 compounds (412 hepatotoxicants and 214 non-hepatotoxicants). The compounds were classified as either hepatotoxic for humans and/or animals or with weak or no evidence of hepatotoxicity. Although its sensitivity and accuracy were close to random (0.46 and 0.56 respectively) and its specificity quite reasonable (0.77), this approach was not designed for screening purposes. Nevertheless, it should be noted that alerts were prioritised based on their applicability to the Pfizer compound collection. Additionally, compounds that showed unambiguous toxicity during *in vitro* screening were not prioritised for *in vivo* studies, and thus were not considered in this study, potentially explaining the very low sensitivity.

In a second approach a set of 244 hepatotoxic compounds was aggregated from the literature and from failed clinical candidates and drugs withdrawn from the market[88]. From these, 74 structural alerts were derived from mechanistic information, of which 56 were related to reactive and toxic metabolites metabolism. The remaining 18 alerts were based on high cut-off similarity queries, as no mechanistic information could be derived. The authors did not evaluate the predictive performance of these structural alerts but deployed them within the VERDI cheminformatics platform from Vertex pharmaceuticals.

In a third approach[89], a diverse set of 951 compounds was compiled through curation of the dataset from Fourches *et al.*[30]. The protein binding potency of each compound was predicted and structural similarity-based clusters of compounds were identified. These categories were then manually curated and related to other well characterised structural alerts. Finally, each alert was thoroughly examined to derive a mechanistic hypothesis for the observed hepatotoxicity. In total 16 structural alerts were characterised. The authors did not validate such alerts on external datasets as their aim was to provide a scheme to identify mechanistically supported structural alerts.

Applying a similar process, Pizzo *et al.* compiled a dataset of 950 compounds of which 510 were hepatotoxicants and identified 13 structural alerts manually and 75 through automatic identification, 11 and 40 of which were respectively associated with hepatotoxicity[90]. The authors then developed an expert-based decision tree based on these structural alerts to predict binary general hepatotoxicity. The model developed was subsequently validated against an external dataset of 101 compounds (69 hepatotoxicants), of which 41% could not be predicted as did not contain any structural alert. Although sensitivity and accuracy were satisfactory for such an approach (0.80 and 0.68 respectively) the model performed poorly in terms of specificity (0.33). Through thorough examination the authors derived a mechanistic hypothesis for the manually derived structural alerts. In addition to the β-lactam substructures, retinoids, oestrogen steroids identified by Hewitt *et al.*[89], the authors characterised N-containing heterocyclic aromatic compounds, sulphonamides, nucleoside analogues, tricyclic antidepressants, aromatic amines, macrolide antibiotics, anti-bacterial agents, cationic amphiphilic drugs to be mostly associated with hepatotoxicity and nitrosourea compounds not to be associated with hepatotoxicity.

Finally, aggregating DILI associated compounds from LiverTox[50] with literature findings, Liu *et al.* performed substructure searches using literature-based structural alerts[91]. Alerts were ranked by their probability of chance occurrence to classify compounds as being hepatotoxic, non-hepatotoxic, or possible hepatotoxic. This led to the identification of 12 statistically relevant alerts that, unfortunately, were not validated on an external set for prospective prediction. In addition to steroids that were already well characterised hepatotoxicants, sulphonamides, hydrazines, arylacetic acids, anilines, sulfinyls, acyclic bivalent sulphurs, acyclic diaryl ketones, halogen atoms bonded to a sp^3 carbon, aminocyclopropyls, aminophenols and phenothiazines were identified as being toxic to the liver.

238 *Table 2: Reported computational models for the prediction of DILI*

year	*ref.*	*endpoint*	*prediction*	*Descriptors*	*data points (positive/negative)*	*methods*	*performance*	*type of data*	*species*
2003	7	general hepatotoxicity	binary	Cerius2 2D molecular descriptors	CV: 382 (149/233) EV: 54 (23/31)	RF	CV: 0.85 Acc, 0.78 Sen, 0.90 Spe EV: 0.81 Acc, 0.70 Sen, 0.90 Spe	*in vivo*	human
2004	92	4 endpoints, general hepatotoxicity	binary	molecular electrostatic field	654	SIMCA	0.52 Acc	*in vitro*	human
2008	33	general hepatotoxicity	binary	radial distribution function	CV: 74 (33/41) EV: 13 (13/0)	LDA ANN one-level DT	CV: 0.86 Acc,0.81 Sen, 0.90 Spe EV: 0.82 Acc CV: 0.78 Acc, 0.75 Sen, 0.80 Spe CV: 0.81 Acc, 0.76 Sen, 0.98 Spe	*in vivo*	human
2009	93	Liver disorders, jaundice and cholestasis, liver enzymes elevation, bile duct disorders	binary	molecular fragment descriptors	CV: 1044 – 1608 EV: 18	4 commercial QSAR programs	CV: 0.32-0.47 Sen, 0.85-0.88 Spe EV: 0.89 Sen	*in vivo*	human
2010	42	AST level, ALT level Composite score	binary	MolconnZ topological descriptors and DRAGON molecular descriptors	CV: 190 (76/114) CV: 210 (84/126) CV: 188 (75/113)	k-NN	CV: 0.74-0.92 Acc, 0.60-0.88 Sen, 0.89-0.96 Spe	*in vivo*	human
2010	30	general hepatotoxicity	binary	ISIDA 2D fragments and DRAGON molecular descriptors	CV: 531 (248/283) EV: 18	SVM	CV: 0.62-0.68 Acc EV: 0.56-0.73 Acc		human rodents non-rodents
2010	40	general hepatotoxicity	binary	extended connectivity fingerprint with counts and bound diameter 6	CV: 295 (158/137) EV: 237 (114/123)	NB	CV:0.58 Acc, 0.53 Sen, 0.64 Spe EV: 0.60 Acc, 0.56 Sen, 0.67 Spe	*in vivo*	human
2010	41	general hepatotoxicity	binary	structural alerts	EV: 626 (412/214)	-	EV: 0.56 Acc, 0.46 Sen, 0.73 Spe	*in vivo*	human
2011	94	13 hepatopathology endpoints	binary	function class fingerprint with counts and bond diameter 6	CV: 22-274 EV: 40-148	NB	CV: 0.93-0.99 Acc EV: 0.60-0.70 Acc	*in vivo*	human
2011	47	general hepatotoxicity	binary	toxicogenomics descriptors	CV: 127 (53/74)	RF, k-NN, SVM	CV: 0.69-0.76 Acc, 0.57-0.67 Sen, 0.77-0.84 Spe	*in vivo*	rats
2011	45	general hepatotoxicity	binary	PaDEL molecular descriptors	CV: 1087 (654/433) EV: 120 (72/48)	SVM, NB, k-NN	CV: 0.64 Acc, 0.64 Sen, 0.63 Spe EV: 0.62 Acc, 0.62 Sen, 0.62 Spe	*in vivo*	human
2012	95	general hepatotoxicity, 3 hepatopathology endpoints	binary	ChemTree augmented atom pairs	CV: 1380 EV: 231-901 3 endpoints EV: 28-539	RF	EV: 0.64-0.81 Acc, 0.58-0.73 Sen, 0.71-0.88 Spe EV: 0.62-1.00 Acc, 0.75-1.00 Sen, 0.60-1.00 Spe		mouse rat
2013	78	general hepatotoxicity	binary	Log P and daily dose	CV: 164 (116/48) EV: 179 (115/64)	Rule of 2	IV: 0.55 Acc, 0.36 Sen, 0.96 Spe EV: 0.51 Acc, 0.29 Sen, 0.91 Spe		human animal
2013	96	general hepatotoxicity	binary	Mold2 chemical descriptors	CV: 197 (81/116) EV: 190-328 (95-214/95-114)	RF	CV: 0.70 Acc, 0.58 Sen, 0.78 Sep EV: 0.62-0.69 Acc, 0.58-0.66 Sen, 0.66-0.72 Spe	*in vivo*	human
2014	97	general hepatotoxicity	binary	CDK, Dragon and MOE molecular descriptors and 8 cellular phenotypes	CV: 292 (156/136)	RF	CV:0.68-0.73 Acc, 0.71-0.73 Sen, 0.64-0.74 Spe	*in vivo*	human
2014	64	general hepatotoxicity	binary	E-dragon molecular descriptors	CV: 872 (436/436) IV: 216 (54/162) EV: 23 (23/0)	SVM	CV: 0.83 Acc IV: 0.82 Acc, 0.87 Sen, 0.81 Spe EV: 0.74 Acc	*in vivo*	human
2015	70	hypertrophy, injury, proliferative lesions	binary	QuikProp physicochemical descriptors, PaDEL fingerprints and *in vitro* bioactivity data	CV: 677 (161/463) (101/463) (99/463)	LDA, NB, SVM, k-NN	CV: 0.62-0.84 BAcc, 0.27-0.77 Sen, 0.85-1.00 Spe	*in vivo*	animal
2015	58	general hepatotoxicity	binary	undirected graph recursive neural networks	CV: 475 (236/239) EV: 198 (114/84)	DL	CV: 0.88 Acc, 0.90 Sen, 0.87 Spe EV: 0.87 Acc, 0.83 Sen, 0.93 Spe	*in vivo*	human
2015	68	general hepatotoxicity	binary	PaDEL molecular descriptors	CV: 201 (136/65) EV: 91 (83/8)	RF	CV: 0.79 Acc, 0.91 Sen, 0.54 Spe EV: 0.87 Acc, 0.90 Sen, 0.63 Spe	*in vivo*	human

year	*ref.*	*endpoint*	*prediction*	*Descriptors*	*data points (positive/negative)*	*methods*	*performance*	*type of data*	*species*
2015	65	7 hepatopathology endpoints, general hepatotoxicity	binary	ISIDA descriptors and *in vivo* endpoints	CV: 414 (41-168) EV: 10 (9/1)	SVM, RF, ANN	QSAR CV: 0.58-71 BAcc Endpoints CV: 0.86-0.87 BAcc Endpoints EV: 0.90 Acc, 0.89 Sen, 1.00 Spe	*in vitro* *in vivo*	human
2015	91	general hepatotoxicity	hepatotoxic non-hepatotoxic possible hepatotoxic	structural alerts	178 185 242	-		*in vivo*	human
2016	74	general hepatotoxicity	binary	FP4 descriptors	CV: 336 (206/130) EV: 84 (51/33)	NB	CV: 0.94 Acc, 0.97 Sen, 0.89 Spe EV: 0.73 Acc, 0.73 Sen, 0.73 Spe	*in vivo*	human
2016	53	21 endpoints	binary	CATS, MOE, MDL, VolSurf+ physicochemical descriptors	CV: 3712 EV: 221-269	SVM with GA	CV: 0.73-0.83 Acc EV: 0.38-0.64 Acc	*in vivo*	animal human
2016	90	general hepatotoxicity	binary	structural alerts	CV: 950 (510/440) EV: 202 (137/65)	expert manual DT	CV: 0.81 Acc, 0.93 Sen, 0.67 Spe EV: 0.68 Acc, 0.80 Sen, 0.33 Spe	*in vivo* *in vitro*	human
2016	76	general hepatotoxicity	multiple scales	Log P, daily dose or Cmax, and formation of metabolites	IV: 192 (124/68)	-	IV: 0.47Acc, 0.38 Sen, 1.00 Spe	*in vivo*	human
2016	98	general hepatotoxicity	binary	CDK, Dragon and Mold2 molecular descriptors, HTS bioactivity data	CV: 233 (67-166)	RF	CV: 0.66-0.73 Acc, 0.62-0.77 Sen, 0.56-0.79 Spe	*in vivo*	mouse
2016	99	general hepatotoxicity	binary	FP4 and MACCS fingerprints	CV: 978 (571/407) IV: 251 (155/96) EV: 88 (59/29)	SVM, NB, k-NN, DT, RF	CV: 0.67-0.82 Acc, 0.92-0.96 Sen, 0.32-0.62 Spe IV: 0.60-0.66, 0.77-0.93 Sen, 0.24-0.34 Spe EV: 0.65-0.75 Acc, 0.81-0.93 Sen, 0.21-0.38 Spe	*in vivo*	human
2017	100	general hepatotoxicity No/Less/Most DILI	binary ternary	Mold2 molecular descriptors	CV: 451 (183/268) EV: 721 (183/270/268)	RF	CV: 0.73 Acc, 0.63 Sen, 0.53 Spe CV: 0.53 Acc	*in vivo* *in vitro* *in vitro*	rat rat human
2017	101	general hepatotoxicity	binary	PubChem fingerprints	CV: 312 (180/132) EV: 398 (224/174)	RF, SVM	CV: 0.73-0.74 Acc EV: 0.61 Acc	*in vivo*	human
2017	102	general hepatotoxicity	binary	Log P and daily dose	IV: 568 (313/255)	Rule of 2	IV: 0.58 Acc, 0.80 Sen, 0.52 Spe	*in vivo*	human
2017	73	general hepatotoxicity	binary	CORAL descriptors	CV: 2029 (662/1367)	Monte Carlo optimization	CV: 0.83-0.87 Acc, 0.71-1.00 Sen, 0.85-0.87 Spe	*in vivo*	human
2017	69	general hepatotoxicity	binary	molecular descriptors	CV: 34023 (64/3339)	NB, Ensemble	CV: 0.78 BAcc, 0.74 Sen, 0.83 Spe CV: 0.60 BAcc, 0.70 Sen, 0.65 Spe		
2017	13	general hepatotoxicity	binary	MACCS public fingerprints, CDK and Mold2 molecular descriptors	CV: 1054 (122/932)	RF	CV: 0.77-0.84 Acc, 0.76-0.88 Sen, 0.73-0.80 Spe	*in vivo*	human
2017	66	general hepatotoxicity	binary	Mold2 descriptors	CV: 192 (127/65) EV: 20 (14/6)	RF	CV: 0.80-0.84 Acc, 0.82-0.84 Sen, 0.70-0.75 Spe EV: 0.90 Acc, 1.00 Sen, 0.67 Spe	*in vivo*	human
2017	103	17 modes of actions	binary	Mold2 descriptors	CV: 222 (155/178) EV: 111	RF	CV: 0.70-0.76 Acc IV: 0.70-0.71 Acc	*in vivo*	human
2018	56	general hepatotoxicity	binary	CDK estate, MACCS, FP4, atom pairs fingerprints	CV: 1241 (683/558) EV: 286 (221/65)	XGBoost, RF, SVM	CV: 0.63-0.70 Acc, 0.66-0.82 Sen, 0.41-0.63 Spe EV: 0.73-0.86 Acc, 0.72-0.89 Sen, 0.42-0.83 Spe	*in vivo*	human
2018	104	general hepatotoxicity	binary	PaDEL molecular descriptors and fingerprints	1731 (980/751) IV: 413 (270/143) EV: 151 (88/63)	SVM, k-NN, NB, DT, RF	IV: 0.62-0.80 Acc, 0.53-0.97 Sen,0.13-0.83 Spe EV: 0.66-0.83 Acc, 0.68-0.93 Sen, 0.54-0.70 Spe	*in vivo*	human
2018	75	general hepatotoxicity	binary	PaDEL descriptors	CV: 712 (444/268)	RF, ANN	CV: 0.80-0.90 Acc, 0.78-0.90 Sen, 0.81-0.90 Spe	*in vivo*	human
2018	105	general hepatotoxicity	binary	PaDEL molecular descriptors	CV: 99 (48/51) EV: 25 (10/15)	k-NN with GA	CV: 0.76 Acc, 0.79 Sen, 0.74 Spe EV: 0.92 Acc, 0.90 Sen, 0.93 Spe	*in vivo*	rats
2018	77	general hepatotoxicity	binary	PaDEL molecular descriptors	CV: 575 (384/191)	DT, k-NN, SVM, ANN	CV: 0.53-0.98 Acc	*in vivo*	human

year	*ref.*	*endpoint*	*prediction*	*Descriptors*	*data points (positive/negative)*	*methods*	*performance*	*type of data*	*species*
2018	52	general hepatotoxicity	binary	maximum daily dose, LogP, Fraction of sp3 carbons	326 (163/163)	Expert manual DT	0.82 Acc, 0.79 Sen, 0.85 Spe	*in vivo*	human
2018	106	hepatocellular hypertrophy	binary	DRAGON molecular descriptors	405 (207/198) EV: 405 (218/187)	ANN, RF, SVM	EV: 0.68-0.76 Acc, 0.58-0.90 Sen, 0.46-0.84 Spe	*in vivo*	rats
2018	72	serum ALT level	binary	DRAGON molecular descriptors	CV: 176 (40/136)	LR	CV: 0.60 Acc, 0.65 Sen, 0.58 Spe EV: 0.60 Sen, 0.40-0.50Acc and Spe	*in vivo*	rats
2018	107	non-neoplasic proliferative lesions inflammatory liver changes degenerative lesions	binary and continuous	Adriana and GRIND2 molecular descriptors	332 (168/164) 258 (164/94) 246 (164/82)	PLS, RF	CV: 0.70 Sen, 0.69 Spe EV: 0.50 Sen, 0.62 Spe CV: 0.44 Sen, 0.84 Spe EV: 0.54 Sen, 0.76 Spe CV: 0.68 Sen, 0.55 Spe EV: 0.67 Sen, 0.59 Spe	*in vitro* *in vivo*	animal
2018	108	general hepatotoxicity	four categories	solubility, *in vitro* permeability, metabolism, dose	EV: 164 (116/48) EV: 192 (124/68)	Rule-based	EV: 0.62-0.72 Acc EV: 0.66-0.78 Acc	*in vivo* *in vitro*	human animal
2019	67	general hepatotoxicity 4 severity degrees 22 adverse events	binary	MOE molecular descriptors	CV: 2513 (1475-1720/1038) CV: 426-1180 (213-590/213-590) CV: 200-1104 (100-552/100-552) EV: 11-16/0	RF	CV: 0.69 Acc, 0.84 Sen, 0.51 Spe CV: 0.70-0.71 Acc, 0.71-0.77 Sen, 0.63-0.70 Spe CV: 0.67-0.78 Acc, 0.65-0.84 Sen, 0.63-0.81 Spe Tiered CV: 0.67 Acc EV: 0.81-0.82 Spe	*in vivo*	human
2019	109	general hepatotoxicity	binary	Marvin molecular descriptors	CV: 1254 (636/618) EV: 204 (125/79)	NB, k-NN, RF, ANN, Ensemble	CV: 0.60-78 Acc, 0.61-0.86 Sen, 0.40-0.76 Spe Ensemble CV: 0.78 Acc, 0.82 Sen, 0.75 Spe Ensemble EV: 0.73 Acc, 0.77 Sen, 0.66 Spe		animal human
2019	71	general hepatotoxicity	binary	PaDEL molecular fingerprints	CV: 1812 (453/1359) IV: 664 (166/498)	ANN, SVM, RF, k-NN, Ensemble	CV: 0.85-0.90 Acc, 0.71-0.86 Sen, 0.82-0.92 Spe IV: 0.82-0.89 Acc, 0.60-0.80 Sen, 0.83-0.93 Spe	*in vivo*	human
2019	83	biliary hyperplasia, fibrosis, and necrosis	binary	transcriptomic data	CV: 2324 (91/2333, 37/2287, 275/2049) EV: 341-376 (20/321, 22/354, 32/326)	DL, RF, SVM	CV: 0.48-0.89 MCC EV: 0.36-0.90 MCC	*in vivo*	rats
2019	110	general hepatotoxicity	binary	PaDEL molecular fingerprints and descriptors	450 (182/268)	LR, SVM, GBT, RF, Ensemble	CV: 0.77 Acc, 0.64 Sen, 0.86 Spe IV: 0.82 Acc, 0.65 Sen, 0.96 Spe	*in vivo*	human
2019	111	general hepatotoxicity	four categories	Log P, daily dose, ionization state, carbon bond saturation and mechanistic assays	CV: 200 (79/56/47/18) IV: 21 EV: 7	Rule of Thumbs	CV: 41-80 Sen, 58-97 Spe	*in vivo*	human
2019	112	general hepatotoxicity	ternary	Log P, Cmax, formation of metabolites and mechanistic assays	96 (33/40/23)	NB	0.63 BAcc Binary: 0.86 Acc, 0.87 Sen, 0.85 Spe,	*in vivo*	human

ANN: artificial neural network, DL: deep learning, DT: decision tree, GA: genetic algorithm, GBT: gradient-boosted trees (of which XGBoost [extremally gradient tree boosting[113]] is an implementation), k-NN: k-nearest neighbours, LDA: latent Dirichlet allocation, LR: logistic regression, NB: naïve Bayes, PLS: partial least squares, RF: random forest, SVM: support vector machine, CV: cross-validation, IV: internal validation, EV: external validation, Acc: accuracy, BAcc: balanced accuracy, Sen: sensitivity, Spe: specificity, MCC: Matthews correlation coefficient

Other studies on the development of quantitative structure-activity relationship (QSAR) models have also focused on the identification of molecular patterns related to hepatotoxicity. Structural fingerprints of compounds (e.g. Kletkota-Roth[114] or extended connectivity fingerprints[115]) have been calculated for a training set. Association of the presence of one pattern with hepatotoxicity was evaluated either based on the feature importance of each bit of such fingerprints or on their frequency. The importance of fingerprint bits has been notably derived from extended connectivity fingerprints with a maximum diameter of 6 (ECFP6) using naïve Bayes models[40,74] and a random forest [56] with 12 different fingerprints. This analysis pointed not only to substructures associated with hepatotoxicity but also those associated with non-hepatotoxic compounds. Frequency focused determination of substructures of interest was performed either by determining the information gain of using such substructures or by using logistic regression, and deriving odds ratios and/or *p*-values associated with these moieties[13,30,45,99,104,106].

The real benefit of using structural alerts is that they may be associated with well characterised mechanisms (e.g. biotransformation to reactive metabolites or alteration in membrane structure integrity, adduction to proteins) and with specific organ level toxicity effects[116]. This reason makes them valuable when determining the toxicity of new drugs and postulating key mechanisms involved. In addition to expert-derived structural alerts, the identification of key substructures associated with DILI is of crucial importance since it allows for further research on, and understanding of, the associated underlying mechanisms.

Nevertheless, a key concept of applying structural alerts is that the absence of a matching alert for a compound is not proof of it not being hepatotoxic[117]. Moreover, the presence of structural alerts should not be seen as a clear indication of the DILI potential of a drug. To emphasise this, Stepan *et al.* retrospectively examined the 200 most prescribed and sold drugs in the US in 2009 and 68 other drugs that had been recalled or were associated with black box warning due to iDILI[118]. Although structural alerts were present in 78%-86% of hepatotoxic drugs, approximately half of the top 200 drugs for 2009 also contained one or more structural alerts, mitigating the use of alerts in for the screening of the toxicity of a compound. According to the authors, "*the major differentiating factor appeared to be the daily dose*", as drugs with high daily doses were mostly associated with toxicity.

Rules Of Thumb

To expand on Stepan *et al.*'s observation about daily dose, few rules of thumb based on two or three molecular features of compounds have been derived. Chen *et al.* identified that from a dataset of 164 US FDA-approved oral medications, a high risk of DILI was associated with lipophilic drugs (Log P ≥ 2) given at high dosage (daily dose ≥ 100 mg; odds ratio 14.05, p-value < 0.001)[78]. This '*rule of two*' was validated using Greene *et al.*'s dataset of 179 oral medications[41]. Of the compounds being positive

for such a rule, 85% were associated with hepatotoxicity. However, this high positive predicted value was associated with very low sensitivity (0.29) but very high specificity (0.91), which overall gives an accuracy (0.51) close to that of a random prediction. When applying this '*rule of two*' to five datasets[29,39,41,48,51], accounting for a total of 1,036 compounds, the authors noticed that the association between toxicity and high lipophilicity was statistically significant for only three of them (those of Chen *et al.*, Greene *et al.* and Zhu *et al.*). Moreover they found that all compounds with a daily dose higher than 100 mg per day were significantly associated with DILI risk[102]. The authors also collected hepatic metabolism information for 398 drugs and observed that drugs, which are more than 50% metabolised in the liver, were more prone to be hepatotoxic (odds ratios between 1.80 and 2.67). Combining significant hepatic metabolism with high daily dose allowed for the correct identification of 78% of hepatotoxic compounds and 60% of non-hepatotoxicants, giving this prediction method an overall accuracy of 0.68. Factoring high lipophilicity with reactive metabolite (RM) formation and high daily dose for a dataset of 192 drugs, the authors were then able to develop a prediction method with a specificity of 1.00 but sensitivity of 0.38[76]. The assessment of the association between daily dose, lipophilicity, RM formation and DILI risk by logistic regression analysis confirmed the significant importance of these features[119] and allowed for the development of a DILI score significantly correlated with the severity of liver injury in human for three different datasets[4,28,41].

Another rule of thumb was derived by Leeson, who investigated the predictivity of physicochemical properties of compounds related to their dose[52]. More specifically, the differences between dose, lipophilicity and the fraction of sp^3 hybridised carbons atoms (Fsp3) in relationship to whether drugs with the most and no DILI concern were acids, bases or neutral (from the Chen *et al.* dataset)[4] were examined. As the mean Fsp3 values of bases, which were enriched in the non-hepatotoxicants class, are greater than for acids[120], the author was able to integrate Fsp3 to the '*rule of two*', yielding accurate predictions for 82% of compounds and with high and balanced sensitivity and specificity (0.79 and 0.85 respectively).

Despite the simplicity of these rules of thumb that have high specificity, their major flaw is that their applicability is limited to the datasets they are built upon[102]. The datasets may have different causality assessment scales to derive DILI annotation which vary from one dataset to the other[121], or reported hepatotoxicity evidence maybe is vague[122,123]. This limitation of the data was stressed by Leeson who identified that among the 155 oral drugs belonging to the top 200 prescribed medications in the US in 2009 that were annotated by Chen *et al*[4], 59% belonged to the Less DILI category, hence questioning the significance of such a class.

QUANTITATIVE STRUCTURE-ACTIVITY AND TOXICITY RELATIONSHIPS

Because the acquisition of some of the parameters mentioned above is only possible from *in vitro* and

in vivo studies QSAR or structure-toxicity relationship-based models have been developed using molecular properties to allow for the early screening of compounds for which no data exist. Examples of experimental properties which may not be available for models include metabolism activity, maximum daily dose or peak concentration in serum after drug administration (Cmax). There are several different types of cheminformatics model: models predicting general hepatotoxicity, histopathological phenotypes (e.g. increase in serum biomarkers, cholangitis) or specific modes of action mediated through protein-ligand interactions.

General hepatotoxicity

Derived from the first phenotypic observations of hepatotoxicity and used to provide a general estimation for compound prioritisation in drug discovery, QSAR models were first built using general binary DILI annotations. For instance, Cheng and Dixon developed one of the first hepatotoxicity QSAR models derived from molecular descriptors, without regard to dose-dependence. In addition to those descriptors, the similarities to the 382 compounds in the training set (149 hepatotoxicants and 233 non-hepatotoxicants) were also used as explanatory variables. Monte Carlo feature selection was applied to reduce the number of descriptors to 25, of which 6 were physicochemical properties. A random forest model was developed and validated on a test set of 54 compounds. Its performance was very encouraging with good accuracy, fair sensitivity and high specificity (0.81, 0.70 and 0.90 respectively). However, such an approach, with such a limited description of the molecular structure and similarity profiles to the training set, did not allow for extrapolation to other compound classes.

Since then, a wide variety of general QSAR models predicting hepatotoxicity have been derived using different types of molecular descriptors, molecular fingerprints and machine learning algorithms (see Table 2). The most recent work predicting general hepatotoxicity solely from molecular descriptors is from He *et al.*[109]. The authors combined a total of 14 datasets for which hepatotoxicity labels originated from animal and cell experiments, clinical reports, drug labels, medical monographs and the scientific literature. In addition, compounds that were classified by fewer than two of eight effective classifiers were discarded, allowing for the creation of a large, balanced and high-quality dataset of 1,254 compounds (636 positives and 638 negatives). Using a set of 85 physicochemical and topological properties an ensemble model based from the eight base classifiers was obtained with high and balanced performance evaluated with 10-fold cross-validation (sensitivity 0.82, specificity 0.75, accuracy 0.78 and balanced accuracy 0.78) and on an external test set of 204 compounds (sensitivity 0.77, specificity 0.66, accuracy 0.73 and balanced accuracy 0.72). To further validate their model to identify non-hepatotoxicants, the authors assembled a dataset of 312 negative compounds. Their classification ensemble model correctly predicted 215 of these compounds, giving a reasonable accuracy of 0.70.

The relevance of building classification models from molecular descriptors alone, in comparison with molecular fingerprints, was questioned by Li *et al.*[104]. The relative performances of k-nearest neighbour

(k-NN), support vector machine (SVM), random forest (RF), naïve Bayesian (NB) and decision tree (DT) models built from seven PaDEL molecular fingerprints[124] and molecular descriptors were compared for a dataset of 980 DILI-positive and 751 DILI-negative compounds. Models based solely on molecular descriptors had the lowest average performance with low accuracy (0.62 to 0.73), specificity (0.13 to 0.70) and AUC (0.0.63 to 0.78). The combination of public MACCS fingerprints in an SVM yielded the best classification performance on an external test set of 88 hepatotoxicants and 63 non-hepatotoxicants (0.83 accuracy, 0.93 sensitivity, 0.68 specificity and 0.88 AUC) despite their limited dimensionality of 166 bits. Only one model, also developed with public MACCS fingerprints but using k-NN, had higher specificity than the previous one (0.70) but lower accuracy, sensitivity and AUC (0.76, 0.81 and 0.82 respectively). This emphasised the usefulness of ensemble models, which was the strategy used by Wu *et al.*[71], who combined four PaDEL molecular fingerprints with k-NN, RF, SVM and artificial neural network (ANN) base classifiers in consensus voting models and also identified the public MACCS fingerprints and SVM-based based classifier to perform well on an external test set of 166 positive and 498 negative compounds (0.75 sensitivity, 0.93 specificity, 0.88 accuracy and 0.70 Matthews correlation coefficient [MCC]). Their consensus models were based on the number of times a compound was predicted to be hepatotoxic by base classifiers. The best performing consensus model, which was that based on three positive predictions out of the 4 base classifiers, was selected (0.77 sensitivity, 0.97 specificity, 0.92 accuracy and 0.78 MCC respectively).

Ai *et al.*[56] adopted the same strategy as Wu *et al.* but filtered out bits of the fingerprints that were correlated and did not apply them to the dataset (e.g. all molecules contain carbon atoms so this information was removed). The five best performing base classifiers in terms of AUC, which interestingly did not include any based on public MACCS fingerprints, were then combined in an ensemble model by averaging their predicted hepatotoxicity probability (0.84 accuracy, 0.87 sensitivity, 0.75 specificity and 0.90 AUC on the external test set).

Wang *et al.*[110] recently combined the Ai *et al.*'s approach with the work of He *et al.* by developing an ensemble model based on the eight PaDEL fingerprints that performed best on their dataset as well as an ensemble model based on seven simple molecular properties (ALogP, molecular weight and numbers of aromatic rings, hydrogen-bond donors, acceptors, rotatable bonds and rings). The five base classifiers used for both these ensemble models were random forest and boosting tree models. The average probabilities for each ensemble were then summed and the weighted average of the two (i.e. 0.7 for fingerprint-based and 0.3 for molecular property-based) were used to classify compounds. The performance of the model was comparable, although slightly lower, than that obtained by Ai *et al.* but specificity was very good (0.82 accuracy, 0.65 sensitivity, 0.96 specificity, 0.80 AUC).

386 ***Phenotypically-focused models***

387 To compensate for the complexity of predicting general hepatotoxicity, models focused on finer
388 phenotypes have been devised. In this sense, Myshkin *et al.* derived an ontology database of hepatotoxic
389 pathology from human and animal publicly available toxicity data[95]. This database was organised by the
390 type of pathology and by organ substructure and function impairment. From this ontology, different
391 toxicity datasets were identified among which were datasets related to liver necrosis, liver weight gain
392 and liver steatosis, comprising of 300, 305 and 172 instances respectively. For each endpoint, random
393 forest QSAR models were derived using augmented atom pairs[125]. The best performing models were
394 then evaluated on external test sets (490, 539 and 478 respectively). Results were encouraging with 0.63,
395 0.74 and 0.60 specificity for liver necrosis, weight gain and liver steatosis respectively, 0.87, 0.86 and
396 0.75 sensitivity, 0.66, 0.76 and 0.62 accuracy and 0.35, 0.51 and 0.23 Matthews correlation coefficient.
397 The authors then characterised the applicability domain of their models based on a Tanimoto distance
398 between compounds in the training and test set. The models were quite sensitive as sensitivity decreased
399 for compounds in the 30-59% compound dissimilarity range. Interestingly, the model based on weight
400 gain was very robust as sensitivity remained above 0.72 for the entire 30-99% range. It is worth
401 mentioning that these three models performed better than a general hepatotoxicity model (0.58
402 sensitivity, 0.71 specificity, 0.64 accuracy and 0.29 Matthews correlation coefficient) which showed a
403 high sensitivity of 0.82 for the 30-39% Tanimoto dissimilarity range, highlighting the relatively high
404 diversity of compounds in the validation set.
405
406 Another work by Takeshita focused on the prediction of alanine transferase (ALT) elevation in rats from
407 repeated-dose toxicity studies[72]. Two logistic regression models, with seven and nine explanatory
408 variables out of an initial 3,636 DRAGON molecular descriptors respectively [126], were derived to
409 classify 176 compounds. Compounds which had either a lowest observed effect level (LOEL) associated
410 to ALT elevation, (40 positives and 136 negatives) or an elevation in ALT at a dose below 1000 mg/kg
411 (23 strong and 153 weak compounds) were included. Because of the imbalance of their datasets, the
412 authors used the SMOTE algorithm[79]. Although classification performance on the training set was
413 limited between toxic and non-toxic compounds (0.65 sensitivity, 0.581 specificity and 0.600 accuracy),
414 the logistic model showed better discrimination between weak and strong compounds (0.78 sensitivity,
415 0.74 specificity and 0.75 accuracy). External validation on a dataset of 59 compounds (23 strong and 36
416 weak compounds) showed decreased performance (0.60 sensitivity and specificity and accuracy
417 between 0.40 and 0.50). Nevertheless, the significant difference between 52 out of a set of 197 molecular
418 descriptors from the training and test sets was observed by the authors, emphasising the need for
419 applicability domain determination.
420
421 Focusing only on *in vivo* hepatocellular hypertrophy in rats, Ambe *et al.* developed deep learning (DL),
422 RF and SVM classification models[106]. The authors collected rat toxicity data following chronic exposure

of more than 27 days from two sources. Models were trained on half of the data of the two datasets (173 and 251 compounds respectively) as well as on half of their combination (405 compounds) and respectively evaluated on their other halves. DL models were clearly overfitted to the data. Their ROC AUC was 1.00 and accuracy, sensitivity, specificity were 0.96 when evaluated on training set, but dropped when the test set was evaluated. However, the DL model based on the combined dataset did not show such behaviour with more equivalent performance between training and test set. This observation could be the combined result of the two-fold increase in the size of the dataset and the reduction of features from 433 and 417 to 385, corresponding to a decrease in dimensionality by 7.7% to 11.1%. The applicability domain of the models was determined using distance in the molecular space to the training set[127] and resulted in 19, 38 and 50 compounds lying outside for two test sets and their combined version respectively. Using a consensus model based on the majority principle, similar predictive performance was achieved. Of the 107 compounds incorrectly predicted by the consensus model, 78 were predicted incorrectly by all three models. These incorrectly predicted compounds were mostly false positives and the authors exemplified the case of flufenoxuron, a benzoylphenyl urea-based insecticide which is not a hepatocellular hypertrophy inducer in rats but is in mouse carcinogenicity studies. This highlighted the need for the development of models in other species not only for better prediction, but also translation between species and understanding of any species-specific mechanisms involved.

Mulliner *et al.* investigated species specific effects by creating hierarchical seven endpoint hepatopathology trees for human and preclinical findings[53]. An additional tree was developed for preclinical findings obtained at doses lower than 500 mg/kg in order to reduce the leverage of high dose toxicants during model development. The endpoints were organised in three different levels: general hepatotoxicity, morphological and clinical chemistry findings, hepatocellular and hepatobiliary injuries. A total of 3,712 compounds were aggregated with overall concordance between human and animal hepatotoxicity of 77%. Individual SVM classification models were developed for each endpoint using a genetic algorithm for feature selection. All human endpoints were reasonably well predicted with accuracies between 0.73 and 0.78 for internal validation. For preclinical endpoints, only general hepatotoxicity could be modelled confidently for toxicity above 500 mg/kg (ROC AUC of 0.73 and lower than 0.67 for others in internal validation). Conversely all endpoints with the exception of hepatobiliary injuries could be modelled for toxicity below such a threshold (accuracies between 0.75 and 0.83 in internal validation). An external validation on 269 proprietary compounds with 14 to 28-day rat study data showed decreased performance for all models (accuracies between 0.38 and 0.64 and ROC AUC between 0.51 and 0.68). The reduction in performance observed between internal and external validation for preclinical data was expected to be similar for human endpoints, more especially when applying these models on early research drug candidates which do not exhibit similar molecular properties as drugs.

A similar work by López-Massaguer *et al.*[107] relied on an ontology to classify compounds for three endpoints as well as predict the LOEL of compounds from the eTOX database[61]. This database was derived from multiple types of publicly available and confidential preclinical data, in multiple species, for various administration routes and for different exposure times. Aggregating rat *in vivo* microscopy and hepatopathology findings, the authors gathered 164, 94 and 82 positive compounds for the three endpoints (i.e. degenerative lesions [DEG], inflammatory liver changes [INF] and non-neoplasic proliferative lesions [PRO]). It is worth noting that the negative compounds that were selected had been tested at concentrations higher than 1000 mg/kg and had no observed treatment-related and liver-related histopathology findings (168, 164 and 164 for DEG, INF and PRO respectively). Sensitivities and specificities of random forest classification models were balanced after both cross and external validation for PRO (0.70 and 0.50 sensitivities and 0.69 and 0.62 specificities at cross and external validation respectively) and DEG (0.68 and 0.67 sensitivities and 0.55 and 0.59 specificities at cross and external validation respectively) while were unbalanced for INF (0.84 and 0. 67 sensitivities and 0.44 and 0.54 specificities at cross and external validation respectively). Partial least square regression models showed poor fit with low goodness-of-fit (ranging from 0.26 to 0.58), poor predictive performance (Q^2 ranging from -0.84 to 0.07) and high standard deviation (ranging from 1 to 2 log units). This work emphasised the possibility of stringent selection of negative compounds as well as aggregation of multiple sources of data containing compounds with different routes of administration and exposure times.

Relying on an hepatopathology-based ontology, as was carried out in the two previous approaches, Liu *et al.* introduced a severity grade in their hierarchical approach[67]. The authors organised their ontology into three levels: level 1 denoted general hepatotoxicity, level 2 corresponded to the severity of the hepatotoxicity and level 3 associated with adverse events (e.g. acute liver failure, cholestasis or AST elevation). A total of 2,017 compounds associated with 403 clinical grade 3 adverse events were collected from SIDER[43,44] and LiverTox[50], amongst other, databases. Individual classification random forest models were built for 22 endpoints. The level 1 classification model, predicting general hepatotoxicity, showed good sensitivity and ROC AUC but low specificity (0.81, 0.75 and 0.50 respectively). Models based on DILI severity showed more balanced sensitivities and specificities (0.70-0.71 and 0.63-0.70 respectively) resulting in comparable or slightly higher ROC AUC (0.75-0.78). Adverse events prediction models showed balanced sensitivity and specificity ranging from 0.65 to 0.83 and from 0.63 to 0.79 respectively, as well as reasonable accuracy (0.67-0.78) and a high ROC AUC (0.71 to 0.87). The 27 models were integrated in a tiered prediction model with high sensitivity (0.82). Because of the limited size of the external validation dataset, adverse events prediction at level 3 was a qualitative assessment of the models. Nevertheless, ticrynafen, which had been withdrawn from the market for association with hepatitis, was predicted by level 3 models to be associated with hepatitis, acute hepatic failure, and hepatocellular injury.

Prediction of Specific Modes of Action

Biological mechanism-focused models have been gaining increasing interest in recent years, under the auspices and needs of the ToxCast and Tox21 initiatives. An example is the work of Wu *et al.*[103], who integrated quantitative high-throughput screening bioassay activity data to develop 17 QSAR models. The profiles of mode of action (MOA) of drugs were predicted with a set of 777 2D molecular descriptors using random forest models. The accuracies of prediction models ranged between 0.63 and 0.67, which was quite encouraging considering the imbalance in the data. Nevertheless, when predicting general hepatotoxicity from the predicted MOA profiles, 5-fold cross-validation on a dataset of 222 compounds (155 hepatotoxicants and 178 non-hepatotoxicants with test set included) gave an accuracy of 0.76 and internal validation on 111 drugs gave accuracy of 0.70. This performance was higher than when using a standard QSAR model (accuracy of 0.66 for cross-validation). Interestingly, the general hepatotoxicity model derived from the top four performing MOA profiles prediction models had slightly higher accuracy on the internal validation set while slightly lower through cross-validation (0.71 and 0.70 respectively). These models could be regarded as underperforming as compared to recent general hepatotoxicity QSAR models, however, it should be noted that only a small number of MOAs were considered in this study with regard to the different mechanisms involved in DILI.

Some other studies on the prediction of MOA profiles have been more focused on specific phenotypes. For instance, an impairment of the function of export pumps and transport proteins in the liver would result in the progress of a cholestatic phenotype. The export pumps comprise the biliary salt export pump (BSEP), the breast cancer resistance protein (BCRP) and the P-glycoprotein (P-gp). The transport proteins are the organic-anion-transporting polypeptides (OATPs). OATPs are members of the solute carrier (SLC) family and transport organic anions. Few models have been developed to predict the inhibition of such proteins. A prospective analysis[82] was carried out to identify OATP1B1 and OATP1B3 inhibitors out of DrugBank[128]. This screening was based on a training dataset of 1,708 compounds (190 inhibitors and 1,518 non-inhibitors) for OATP1B1 and of 1,725 compounds (124 inhibitors and 1,601 non-inhibitors) for OATP1B3, respectively. An external test set containing 201 compounds for OATP1B1 (64 inhibitors and 137 non-inhibitors) and 209 compounds for OATP1B3 (40 inhibitors and 169 non-inhibitors) was used to assess the validity of the model along with 5-fold and 10-fold cross-validation. Two random forests and four support vector machine classifiers, using MetaCost[87] as metaclassifier to deal with the imbalance of the dataset, were generated for each transporter. As the performance of the models was relatively equivalent – accuracy values and ROC AUC for the test set in the range of 0.81−0.86 and of 0.81−0.92, respectively – a consensus scoring approach was used, summing up the prediction scores of each classification model. The screening of DrugBank (6,279 compounds) resulted in the identification and biological testing of the 9 compounds with highest predicted probability of being OATPB1 and O1TPB3 dual inhibitors and 1 selective inhibitor of OATP1B3. Only the latter was incorrectly predicted, yielding an accuracy of 90% for OATP1B1 and

80% for OATP1B3, respectively.

To compare the prediction of an inhibitory effect of transport proteins to a phenotypic readout, the relative performance of meta classifiers on unbalanced datasets was studied for OATP1B1 and OATP1B3 inhibition, human cholestasis and animal cholestasis based on molecular descriptors[85,129]. Although imbalance ratios between negatives and positives ranged from 2:1 to 20:1, the balanced accuracies of models with sensitivity higher than 0.5 ranged from 0.67 to 0.83 for OATPB1, 0.63 to 0.86 for OATP1B3 and 0.64 to 0.78 for human cholestasis on test set and from 0.53 to 0.65 for animal cholestasis. This emphasised the difficulty in predicting a phenotypic outcome solely from compound structure.

Other work focused on the prediction of BSEP and MRP4 inhibition from both statistical and structure-based approaches[130]. In this study, 57 and 171 compounds along with inhibitory effect on MRP4 and BSEP were gathered respectively. Bayesian models were trained on simple molecular descriptors and either extended-connectivity fingerprints maximum diameter 6 (ECFP6) or functional-class fingerprints maximum diameter 6 (FCFP6). For MRP4, although the models performed well in terms of specificity, they did not show high sensitivity. Nevertheless, the MRP4 pharmacophore model built on 9 compounds was able to correctly classify 30 of the 42 actives in the test set and 22 of the 35 inactives, leading to a sensitivity of 0.71 and specificity of 0.63. The BSEP inhibition prediction model showed more balanced and higher performance (sensitivity of 0.82 and 0.77, specificity of 0.77 and 0.84 respectively) but the pharmacophore model had a higher selectivity whilst poor specificity of 0.37. The lower performance of the MRP4 classification model was probably due to the 3:1 ratio between active and inactive compounds in the training dataset and to the small size of the dataset comprising only 86 compounds. This work emphasised not only the usefulness of structure-based modelling when it comes to the prediction of inhibitory effects of compounds but also the requirement for well-balanced datasets.

This difficulty to predict a phenotypic outcome of a compound using an imbalanced dataset was tackled using metaclassifiers and considering the predicted inhibitory effect of compounds on transport proteins as descriptors[84]. Cholestasis-focused data were aggregated by mining and manually curating the literature for human drug-induced cholestasis. A total of 578 compounds were identified, of which 131 were cholestasis positives and 447 were DILI negatives. A k-NN classifier with MetaCost metaclassifier for data imbalance correction was generated and evaluated through both 10-fold cross-validation and external testing on a dataset covering multiple levels of hepatotoxicity and including hepatobiliary injury[53]. Inclusion of BSEP, BCRP, P-glycoprotein, and OATP1B1 and OATP1B3 inhibition predictions increased accuracy (0.66 to 0.70) and ROC AUC (0.66 to 0.73) of the model through 10-fold cross validation but decreased for the test set (0.61 to 0.56 and 0.62 to 0.58 respectively). The authors speculated that this was the result of a different class assignments between the training and test

sets and argued that almost 20% of the compounds in the external validation set had contradictory labels with the training set (71 out of 419 shared compounds). Nevertheless, the authors showed that accuracy and specificity reach their peak only after the inclusion of BSEP predictions, but that when only using BSEP predictions, the model showed a slight increase in accuracy and specificity of the model but decreased sensitivity. This suggested that BSEP inhibition conveys most, but not all, of the relevant information when modelling cholestasis.

An effort to merge multiple publicly available datasets was undertaken to apply the models obtained to other datasets and investigate how export pump and transporter inhibition correlate to general hepatotoxicity[59]. In this work, the authors gathered nine previously published datasets for model training (966 compounds) and three datasets for validation (996 compounds). Three random forests classifiers were built using two sets of molecular descriptors to predict transporter inhibition[82,131,132]. Accuracy and ROC AUC of the models ranged from 0.57 to 0.69 and from 0.59 to 0.73 respectively in spite of the heterogeneity of such a dataset, ranging from *in vitro* cell-based assay readouts to FDA reports and post-marketing safety data. Nevertheless, the introduction of BSEP, BCRP, P-glycoprotein, and OATP1B1 and OATP1B3 inhibition binary prediction as descriptors slightly decreased the model performance. The authors argued that this could be the result of mispredictions of such transporter inhibition models resulting in noise added to the feature matrix and that the inhibition of only one transporter would not alter the function of hepatocytes. With regards to such possible misclassifications, the use of a hard threshold at 10 μM to classify a compound as being an inhibitor can lead to misclassification of compounds with IC50 around such a threshold, thus artificially lowering the performance of the model. Additionally, such a threshold is not in accordance with the 300 μM value that was suggested to be used for BSEP inhibition[133]. QSAR models modelling BSEP inhibition based on the latter threshold showed very good performance[134,135]. Finally, the endpoint to be predicted denotes general phenotypic hepatotoxicity and correlates only with transporter inhibition which is associated mostly with cholestasis.

It should also be noted that the BSEP, BCRP, P-glycoprotein, and OATP1B1 and OATP1B3 do not represent the entirety of transporters. One could also cite the canicular and basolateral multidrug resistance-associated proteins (MRP1 to MRP6), the organic solute transporters (OSTα/OSTβ), the multidrug and toxin extrusion transporter 1 (MATE1), the ATP-binding cassette subfamily G member 5/8 (ABCG5/G8), the multidrug resistance protein 3 (MDR3), the ATPase-aminophospholipid transporter (ATP8B1), the sodium taurocolate co-transporting polypeptide (NTCP), the organic cation transporters 1 and 3 (OCT1/3), the organic anion transporters 2 and 7 (OAT2/7) and other organic anion transporting polypeptides (e.g. OATP2B1)[136]. However, to date, very few inhibition data have been collected for these targets, making such a modelling exercise rather difficult if not unfeasible.

608 Finally, Peng et *al.* developed MOA prediction models in the context of steatosis[137]. Data from 24 *in*
609 *vitro* HTS assays from the ToxCast program were compiled. The agonistic and/or antagonistic activity
610 toward six transcription factors (namely the pregnane X receptor [PXR], liver X receptor [LXR], aryl
611 hydrocarbon receptor [AhR], nuclear factor (erythroid-derived 2)-like 2 [Nrf2], PPARα and PPARγ)
612 were modelled using DRAGON molecular descriptors and random forest models. For each MOA, four
613 models were developed based on different strategies in feature selection and class balancing (i.e.
614 majority class undersampling or balanced bagging) and integrated in a consensus model. External
615 validation of the consensus models showed very good performance for all MOAs (accuracy between
616 0.74 and 0.96) but for agonistic activity on PPARγ (accuracy of 0.66) for compounds in the applicability
617 domains. A second validation was carried out by screening 90 chemicals with *in vitro* steatosis data (six
618 positives, 84 negatives) without experimental data for the molecular initiating events (MIE) endpoints
619 considered and gave perfect sensitivity and AUC of 0.72. This exemplified how modelling the MIE can
620 be successfully integrated in a virtual screening strategy for identifying chemicals causing hepatic
621 steatosis.
622

623 **GENERAL DISCUSSION**

624 Predicting DILI is a vital task, but is fraught with difficulties and complexities brought about from the
625 data available to model, the number and varieties of phenotypic endpoint and mechanisms and the
626 requirements of the end user. In the last decade, many QSAR and few rule-of-thumb models have been
627 developed for the prediction of DILI with the majority of them focused on classification of compounds
628 based on general hepatotoxicity annotation (Table 2).The good performance of models that have been
629 developed is very encouraging, highlighting that machine learning methods are able to cope with
630 complexities of the datasets, even though the data is inherently variable, limited in size and imbalanced.
631 This is even more exciting considering that hepatotoxicity is an umbrella term for many different and
632 complex phenotypes that are the integrated result of various mechanisms, and in spite of the paucity of
633 phenotypically- and mechanistically-based large datasets. It is worth noting that only one regression
634 model correlating to the severity of clinical outcome has been published so far[76]. The same applies to
635 multinomial classification modelling: only one three-level DILI classification model has been
636 published[100]. Nevertheless, as no golden standard for DILI annotation has been established, each
637 annotation uses its own criteria and sources to label compounds[102], leading to contradictory
638 hepatotoxicity labelling of compounds by different authors, thus making the integration of multiple
639 datasets a difficult endeavour[59,121]. This stresses the requirement for sensitive biomarkers able to
640 accurately differentiate medical symptoms of DILI. However, the downside of using more complex
641 machine learning algorithms is that they lack transparency and accountability.
642

643 Additionally, differences in molecular similarity among datasets[77,99,104,109] as well as their evaluation

with different metrics makes fair comparison between models a challenge[138]. Among molecular descriptors, there seems to be a growing trend in using molecular fingerprints only, rather than relying on physicochemical or topological descriptors, although simple rules of thumbs have been devised from them. To date only one study has used graph-based molecular structural encoding, thus avoiding the molecular descriptor calculation and selection step, combined with deep learning algorithms[58]. Some other studies have focused on matched molecular pairs – i.e. molecules that are structurally very similar – with opposing hepatotoxicity annotations[30,42,45,47].

Standard physicochemical and topological descriptors, as well as substructure-based fingerprints in QSAR models (structural alerts excluded), are poor predictors of the reactivity of the molecules and its relationship to the metabolism and hence generally do not perform well to predict DILI. In addition, the development of prediction models able to correctly predict toxicity cliffs (i.e. where a very small change in the structure of a molecule can alter activity enormously) is a challenging field. Tackling toxicity cliffs both through better data compilation and more detailed structure evaluation would definitely help better understanding the mechanisms underlying DILI. Hybrid models integrating molecular descriptors with *in vitro* data, whether being transcriptomics[47], cell-imaging[97] or bioactivity data[65,70,137], have also been developed to enrich the information content and interpretability of the models but with rather limited predictive performance. Only a few models have included *in vivo* pharmacokinetic processes, such as absorption and metabolism inhibition of CYP450 proteins, the formation of GSH adducts and protein covalent-binding data[48,139]. Additionally, models focused on the determination of MIE and MOA show very good performance and are of critical importance for better understanding of DILI mechanisms. Yet, it is striking that no ensemble read-across approach, combining systems biology network analysis for the prediction of molecular targets[140], MIE or MOA along with transcriptomics[141,142], cell-imaging and metabolomics, has been devised to this date. Such an approach, similar to the DILIsym[143] systems toxicology strategy, could address the limitations of QSAR[144] such as the modelling of chemical mixtures or inorganic compounds (e.g. cisplatin) as well as enhance models developed this far with the prediction of the exposure. Furthermore, computational structure-based mechanistic hypothesising is very limited by the lack of three-dimensional structures of proteins at stake. Additionally, since dose is an important predictor for DILI, the prediction of the toxicological point of departure[145] (POD) is challenge to be addressed. Finally, the most difficult challenge is to address inter-species variability, and the concordance between human and animal toxicity[30,100,146] that initiatives, such as the eTRANSAFE consortium, focus on.

ACKNOWLEDGEMENTS

The authors thank Peter Hartog for his excellent assistance in generating Figure 1.

FUNDING

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365 ("eTRANSAFE") receiving support from the European Union's Horizon 2020 research and innovation programme and EFPIA and was part of the EU-ToxRisk project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 681002.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

(1) Holt, M.; Ju, C. Drug-Induced Liver Injury. In *Annals of Internal Medicine*; 2010; Vol. 137, pp 3–27. https://doi.org/10.1007/978-3-642-00663-0_1.

(2) Siramshetty, V. B.; Nickel, J.; Omieczynski, C.; Gohlke, B. O.; Drwal, M. N.; Preissner, R. WITHDRAWN - A Resource for Withdrawn and Discontinued Drugs. *Nucleic Acids Res.* **2016**, *44* (D1), D1080–D1086. https://doi.org/10.1093/nar/gkv1192.

(3) Fung, M.; Thornton, A.; Mybeck, K.; wu, J. H. H.; Hornbuckle, K.; Muniz, E. Evaluation of the Characteristics of Safety Withdrawal of Prescription Drugs from Worldwide Pharmaceutical Markets-1960 to 1999. *Ther. Innov. Regul. Sci.* **2001**, *35* (1), 293–317. https://doi.org/10.1177/009286150103500134.

(4) Chen, M.; Vijay, V.; Shi, Q.; Liu, Z.; Fang, H.; Tong, W. FDA-Approved Drug Labeling for the Study of Drug-Induced Liver Injury. *Drug Discov. Today* **2011**, *16* (15–16), 697–703. https://doi.org/10.1016/j.drudis.2011.05.007.

(5) Babai, S.; Auclert, L.; Le-Louët, H. Safety Data and Withdrawal of Hepatotoxic Drugs. *Therapie* **2018**. https://doi.org/10.1016/j.therap.2018.02.004.

(6) van Tonder, J. J.; Steenkamp, V.; Gulumi, M. Pre-Clinical Assessment of the Potential Intrinsic Hepatotoxicity of Candidate Drugs. In *New Insights into Toxicity and Drug Testing*; InTech, 2013. https://doi.org/10.5772/54792.

(7) Cheng, A.; Dixon, S. L. In Silico Models for the Prediction of Dose-Dependent Human Hepatotoxicity. *J. Comput. Aided. Mol. Des.* **2003**, *17* (12), 811–823. https://doi.org/10.1023/B:JCAM.0000021834.50768.c6.

(8) Devarbhavi, H. An Update on Drug-Induced Liver Injury. *J. Clin. Exp. Hepatol.* **2012**, *2* (3), 247–259. https://doi.org/10.1016/j.jceh.2012.05.002.

(9) Zimmerman, H. J. Drug-Induced Liver Disease. *Clin. Liver Dis.* **2000**, *4* (1), 73–96, vi. https://doi.org/10.1016/S1089-3261(05)70097-0.

(10) Senior, J. R. What Is Idiosyncratic Hepatotoxicity? What Is It Not? *Hepatology* **2008**, *47* (6), 1813–1815. https://doi.org/10.1002/hep.22332.

(11) Kaplowitz, N. Idiosyncratic Drug Hepatotoxicity. *Nat. Rev. Drug Discov.* **2005**, *4* (6), 489–499. https://doi.org/10.1038/nrd1750.

(12) George, N.; Chen, M.; Yuen, N.; Hunt, C. M.; Suzuki, A. Interplay of Gender, Age and Drug Properties on Reporting Frequency of Drug-Induced Liver Injury. *Regul. Toxicol. Pharmacol.* **2018**.

(13) Zhu, X.-W.; Li, S.-J. In Silico Prediction of Drug-Induced Liver Injury Based on Adverse Drug Reaction Reports. *Toxicol. Sci.* **2017**, *158* (2), 391–400. https://doi.org/10.1093/toxsci/kfx099.

(14) Mosedale, M.; Watkins, P. B. Drug-Induced Liver Injury: Advances in Mechanistic Understanding That Will Inform Risk Management. *Clin. Pharmacol. Ther.* **2017**, *101* (4), 469–480. https://doi.org/10.1002/cpt.564.

(15) Alempijevic, T.; Zec, S.; Milosavljevic, T. Drug-Induced Liver Injury: Do We Know Everything? *World Journal of Hepatology*. 2017. https://doi.org/10.4254/wjh.v9.i10.491.

(16) Fraser, K.; Bruckner, D. M.; Dordick, J. S. Advancing Predictive Hepatotoxicity at the Intersection of Experimental, in Silico, and Artificial Intelligence Technologies. *Chem. Res. Toxicol.* **2018**, *31* (6), 412–430. https://doi.org/10.1021/acs.chemrestox.8b00054.

(17) Noureddin, N.; Kaplowitz, N. Overview of Mechanisms of Drug-Induced Liver Injury (DILI) and Key Challenges in DILI Research; 2018; Vol. 1990, pp 3–18. https://doi.org/10.1007/978-1-4939-7677-5_1.

(18) Vinken, M. Adverse Outcome Pathways and Drug-Induced Liver Injury Testing. *Chem. Res. Toxicol.* **2015**, *28* (7), 1391–1397. https://doi.org/10.1021/acs.chemrestox.5b00208.

(19) O'Connell, T. M.; Watkins, P. B. The Application of Metabonomics to Predict Drug-Induced Liver Injury. *Clin. Pharmacol. Ther.* **2010**, *88* (3), 394–399. https://doi.org/10.1038/clpt.2010.151.

(20) Przybylak, K. R.; Cronin, M. T. D. In Silico Models for Drug-Induced Liver Injury -- Current Status. *Expert Opin. Drug Metab. Toxicol.* **2012**, *8* (2), 201–217. https://doi.org/10.1517/17425255.2012.648613.

(21) Chan, R.; Benet, L. Z. *Evaluation of the Relevance of DILI Predictive Hypotheses in Early Drug Development: Review of: In Vitro Methodologies vs. BDDCS Classification*; 2018; Vol. 7. https://doi.org/10.1039/c8tx00016f.

(22) Kuijper, I. A.; Yang, H.; Van De Water, B.; Beltman, J. B. Unraveling Cellular Pathways Contributing to Drug-Induced Liver Injury by Dynamical Modeling. *Expert Opin. Drug Metab. Toxicol.* **2017**, *13* (1), 5–17. https://doi.org/10.1080/17425255.2017.1234607.

(23) Bhattacharya, S.; Shoda, L. K. M.; Zhang, Q.; Woods, C. G.; Howell, B. A.; Siler, S. Q.; Woodhead, J. L.; Yang, Y.; McMullen, P.; Watkins, P. B.; et al. Modeling Drug- and Chemical-Induced Hepatotoxicity with Systems Biology Approaches. *Front. Physiol.* **2012**, *3 DEC* (December), 1–18. https://doi.org/10.3389/fphys.2012.00462.

(24) Roth, R. A.; Ganey, P. E. Intrinsic versus Idiosyncratic Drug-Induced Hepatotoxicity--Two Villains or One? *J. Pharmacol. Exp. Ther.* **2010**, *332* (3), 692–697. https://doi.org/10.1124/jpet.109.162651.

(25) Corsini, A.; Ganey, P.; Ju, C.; Kaplowitz, N.; Pessayre, D.; Roth, R.; Watkins, P. B.; Albassam, M.; Liu, B.; Stancic, S.; et al. Current Challenges and Controversies in Drug-Induced Liver Injury. *Drug Saf.* **2012**, *35* (12), 1099–1117. https://doi.org/10.2165/11632970-000000000-00000.

(26) Ludwig, J.; Axelsen, R. Drug Effects on the Liver - An Updated Tabular Compilation of Drugs and Drug-Related Hepatic Diseases. *Dig. Dis. Sci.* **1983**, *28* (7), 651–666. https://doi.org/10.1007/BF01299927.

(27) O'Brien, P. J.; Irwin, W.; Diaz, D.; Howard-Cofield, E.; Krejsa, C. M.; Slaughter, M. R.; Gao, B.; Kaludercic, N.; Angeline, A.; Bernardi, P.; et al. High Concordance of Drug-Induced Human Hepatotoxicity with in Vitro Cytotoxicity Measured in a Novel Cell-Based Model Using High Content Screening. *Arch. Toxicol.* **2006**, *80* (9), 580–604. https://doi.org/10.1007/s00204-006-0091-3.

(28) Suzuki, A.; Andrade, R. J.; Bjornsson, E.; Lucena, M. I.; Lee, W. M.; Yuen, N. A.; Hunt, C. M.; Freston, J. W. Drugs Associated with Hepatotoxicity and Their Reporting Frequency of Liver Adverse Events in VigiBase™. *Drug Saf.* **2010**, *33* (6), 503–522. https://doi.org/10.2165/11535340-000000000-00000.

(29) Chen, M.; Suzuki, A.; Thakkar, S.; Yu, K.; Hu, C.; Tong, W. DILIrank: The Largest Reference Drug List Ranked by the Risk for Developing Drug-Induced Liver Injury in Humans. *Drug Discov. Today* **2016**, *21* (4), 648–653. https://doi.org/10.1016/j.drudis.2016.02.015.

(30) Fourches, D.; Barnes, J. C.; Day, N. C.; Bradley, P.; Reed, J. Z.; Tropsha, A. Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species. *Chem. Res. Toxicol.* **2010**, *23* (1), 171–183. https://doi.org/10.1021/tx900326k.

(31) Kotsampasakou, E.; Montanari, F.; Ecker, G. F. Predicting Drug-Induced Liver Injury: The Importance of Data Curation. *Toxicology* **2017**, *389* (June), 139–145. https://doi.org/10.1016/j.tox.2017.06.003.

(32) Luo, G.; Shen, Y.; Yang, L.; Lu, A.; Xiang, Z. A Review of Drug-Induced Liver Injury Databases. *Arch. Toxicol.* **2017**, *91* (9), 3039–3049. https://doi.org/10.1007/s00204-017-2024-8.

(33) Cruz-Monteagudo, M.; Cordeiro, M. N. D. S.; Borges, F. Computational Chemistry Approach for the Early Detection of Drug-Induced Idiosyncratic Liver Toxicity. *J. Comput. Chem.* **2008**, *29* (4), 533–549. https://doi.org/10.1002/jcc.20812.

(34) Zimmerman, H. . *Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver.*, 2nd ed.; Lippincott Williams and Wilkins: Philadelphia, 1999.

(35) Guo, J. J.; Wigle, P. R.; Lammers, K.; Vu, O. Comparison of Potentially Hepatotoxic Drugs among Major US Drug Compendia. *Res. Soc. Adm. Pharm.* **2005**, *1* (3), 460–479. https://doi.org/10.1016/j.sapharm.2005.06.005.

(36) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29* (8), 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135.

(37) EPA's National Center for Computational Toxicology. ToxCast Database (invitroDB) https://epa.figshare.com/articles/ToxCast_Database_invitroDB_/6062623. https://doi.org/10.23645/epacomptox.6062623.v2.

(38) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* **2007**, *95* (1), 5–12. https://doi.org/10.1093/toxsci/kfl103.

(39) Xu, J. J.; Henstock, P. V.; Dunn, M. C.; Smith, A. R.; Chabot, J. R.; de Graaf, D. Cellular Imaging Predictions of Clinical Drug-Induced Liver Injury. *Toxicol. Sci.* **2008**, *105* (1), 97–105. https://doi.org/10.1093/toxsci/kfn109.

(40) Ekins, S.; Williams, A. J.; Xu, J. J. A Predictive Ligand-Based Bayesian Model for Human Drug-Induced Liver Injury. *Drug Metab. Dispos.* **2010**, *38* (12), 2302–2308. https://doi.org/10.1124/dmd.110.035113.

(41) Greene, N.; Fisk, L.; Naven, R. T.; Note, R. R.; Patel, M. L.; Pelletier, D. J. Developing Structure-Activity Relationships for the Prediction of Hepatotoxicity. *Chem. Res. Toxicol.* **2010**, *23* (7), 1215–1222. https://doi.org/10.1021/tx1000865.

(42) Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling Liver-Related Adverse Effects of Drugs Using Knearest Neighbor Quantitative Structure-Activity Relationship Method. *Chem. Res. Toxicol.* **2010**, *23* (4), 724–732. https://doi.org/10.1021/tx900451r.

(43) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A Side Effect Resource to Capture Phenotypic Effects of Drugs. *Mol. Syst. Biol.* **2010**, *6* (343), 1–6. https://doi.org/10.1038/msb.2009.98.

(44) Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The SIDER Database of Drugs and Side Effects. *Nucleic Acids Res.* **2016**, *44* (D1), D1075–D1079. https://doi.org/10.1093/nar/gkv1075.

(45) Liew, C. Y.; Lim, Y. C.; Yap, C. W. Mixed Learning Algorithms and Features Ensemble in Hepatotoxicity Prediction. *J. Comput. Aided. Mol. Des.* **2011**, *25* (9), 855–871. https://doi.org/10.1007/s10822-011-9468-3.

(46) Liu, Z.; Shi, Q.; Ding, D.; Kelly, R.; Fang, H.; Tong, W.

Translating Clinical Findings into Knowledge in Drug Safety Evaluation - Drug Induced Liver Injury Prediction System (DILIps). *PLoS Comput. Biol.* **2011**.

(47) Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; et al. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chem. Res. Toxicol.* **2011**, *24* (8), 1251–1262. https://doi.org/10.1021/tx200148a.

(48) Sakatis, M. Z.; Reese, M. J.; Harrell, A. W.; Taylor, M. A.; Baines, I. A.; Chen, L.; Bloomer, J. C.; Yang, E. Y.; Ellens, H. M.; Ambroso, J. L.; et al. Preclinical Strategy to Reduce Clinical Hepatotoxicity Using in Vitro Bioactivation Data for >200 Compounds. *Chem. Res. Toxicol.* **2012**, *25* (10), 2067–2082. https://doi.org/10.1021/tx300075j.

(49) Chen, M.; Zhang, J.; Wang, Y.; Liu, Z.; Kelly, R.; Zhou, G.; Fang, H.; Borlak, J.; Tong, W. The Liver Toxicity Knowledge Base: A Systems Approach to a Complex End Point. *Clin. Pharmacol. Ther.* **2013**.

(50) Hoofnagle, J. H. LiverTox: A Website on Drug-Induced Liver Injury. *Drug-Induced Liver Dis.* **2013**, 725–732. https://doi.org/10.1016/B978-0-12-387817-5.00040-6.

(51) Zhu, X.; Kruhlak, N. L. Construction and Analysis of a Human Hepatotoxicity Database Suitable for QSAR Modeling Using Post-Market Safety Data. *Toxicology* **2014**, *321* (1), 62–72. https://doi.org/10.1016/j.tox.2014.03.009.

(52) Leeson, P. D. Impact of Physicochemical Properties on Dose and Hepatotoxicity of Oral Drugs. *Chem. Res. Toxicol.* **2018**, *31* (6), 494–505. https://doi.org/10.1021/acs.chemrestox.8b00044.

(53) Mulliner, D.; Schmidt, F.; Stolte, M.; Spirkl, H. P.; Czich, A.; Amberg, A. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope. *Chem. Res. Toxicol.* **2016**, *29* (5), 757–767. https://doi.org/10.1021/acs.chemrestox.5b00465.

(54) Huang, R. A Quantitative High-Throughput Screening Data Analysis Pipeline for Activity Profiling; Zhu, H., Xia, M., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2016; Vol. 1473, pp 111–122. https://doi.org/10.1007/978-1-4939-6346-1_12.

(55) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K Chemical Profiles for in Vivo Toxicity Prediction and Mechanism Characterization. *Nat. Commun.* **2016**. https://doi.org/10.1038/ncomms10425.

(56) Ai, H.; Chen, W.; Zhang, L.; Huang, L.; Yin, Z.; Hu, H.; Zhao, Q.; Zhao, J.; Liu, H. Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicol. Sci.* **2018**, No. August, 1–8. https://doi.org/10.1093/toxsci/kfy121.

(57) Copple, I. M.; den Hollander, W.; Callegaro, G.; Mutter, F. E.; Maggs, J. L.; Schofield, A. L.; Rainbow, L.; Fang, Y.; Sutherland, J. J.; Ellis, E. C.; et al. Characterisation of the NRF2 Transcriptional Network and Its Response to Chemical Insult in Primary Human Hepatocytes: Implications for Prediction of Drug-Induced Liver Injury. *Arch. Toxicol.* **2019**, *93* (2), 385–399. https://doi.org/10.1007/s00204-018-2354-1.

(58) Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep Learning for Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2015**, *55* (10), 2085–2093. https://doi.org/10.1021/acs.jcim.5b00238.

(59) Kotsampasakou, E.; Montanari, F.; Ecker, G. F. Predicting Drug-Induced Liver Injury: The Importance of Data Curation. *Toxicology* **2017**, *389*, 139–145. https://doi.org/10.1016/j.tox.2017.06.003.

(60) Steger-Hartmann, T.; Pognan, F.; Sanz, F.; Diaz, C. A. In Silico Prediction of in Vivo Toxicities (ETox)—The Innovative Medicines Initiative Approach. *Toxicol. Lett.* **2009**, *189* (2009), S258. https://doi.org/10.1016/j.toxlet.2009.06.374.

(61) Cases, M.; Briggs, K.; Steger-Hartmann, T.; Pognan, F.; Marc, P.; Kleinöder, T.; Schwab, C. H.; Pastor, M.; Wichard, J.; Sanz, F. The ETOX Data-Sharing Project to Advance in Silico Drug-Induced Toxicity Prediction. *Int. J. Mol. Sci.* **2014**, *15* (11), 21136–21154. https://doi.org/10.3390/ijms151121136.

(62) Pognan, F. Detection, Elimination, Mitigation, and Prediction of Drug-Induced Liver Injury in Drug Discovery. In *Drug-Induced Liver Toxicity, Methods in Pharmacology and Toxicology*; Chen, M., Will, Y., Eds.; Springer, 2018; pp 21–43. https://doi.org/10.1007/978-1-4939-7677-5_2.

(63) Sanz, F.; Pognan, F.; Steger-Hartmann, T.; Díaz, C. Legacy Data Sharing to Improve Drug Safety Assessment: The ETOX Project. *Nature Reviews Drug Discovery*. October 13, 2017, pp 811–812. https://doi.org/10.1038/nrd.2017.177.

(64) Jiang, L.; He, Y.; Zhang, Y. Prediction of Hepatotoxicity of Traditional Chinese Medicine Compounds by Support Vector Machine Approach. *Int. Conf. Syst. Biol. ISB* **2014**, No. 81173522, 27–30. https://doi.org/10.1109/ISB.2014.6990426.

(65) Muller, C.; Pekthong, D.; Alexandre, E.; Marcou, G.; Horvath, D.; Richert, L.; Varnek, A. Prediction of Drug Induced Liver Injury Using Molecular and Biological Descriptors. *Comb. Chem. High Throughput Screen.* **2015**, *18* (3), 315–322. https://doi.org/10.2174/1386207318666150305144650.

(66) Zhao, P.; Liu, B.; Wang, C. Hepatotoxicity Evaluation of Traditional Chinese Medicines Using a Computational Molecular Model. *Clin. Toxicol.* **2017**, *55* (9), 996–1000. https://doi.org/10.1080/15563650.2017.1333123.

(67) Liu, L.; Fu, L.; Zhang, J. W.; Wei, H.; Ye, W. L.; Deng, Z. K.; Zhang, L.; Cheng, Y.; Ouyang, D.; Cao, Q.; et al. Three-Level Hepatotoxicity Prediction System Based on Adverse Hepatic Effects. *Mol. Pharm.* **2019**, *16* (1), 393–408. https://doi.org/10.1021/acs.molpharmaceut.8b01048.

(68) Huang, S. H.; Tung, C. W.; Fülöp, F.; Li, J. H. Developing a QSAR Model for Hepatotoxicity Screening of the Active Compounds in Traditional Chinese Medicines. *Food Chem. Toxicol.* **2015**, *78*, 71–77. https://doi.org/10.1016/j.fct.2015.01.020.

(69) Lu, Y.; Liu, L.; Lu, D.; Cai, Y.; Zheng, M.; Luo, X.; Jiang, H.; Chen, K. Predicting Hepatotoxicity of Drug Metabolites Via an Ensemble Approach Based on Support Vector Machine. *Comb. Chem. High Throughput Screen.* **2017**, *20* (10), 839–849. https://doi.org/10.2174/1386207320666171121113255.

(70) Liu, J.; Mansouri, K.; Judson, R. S.; Martin, M. T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R. S.; Shah, I. Predicting Hepatotoxicity Using ToxCast in Vitro Bioactivity and Chemical Structure. *Chem. Res. Toxicol.* **2015**, *28* (4), 738–751. https://doi.org/10.1021/tx500501h.

(71) Wu, Q.; Cai, C.; Guo, P.; Chen, M.; Wu, X.; Zhou, J.; Luo, Y.; Zou, Y.; Liu, A.; Wang, Q.; et al. In Silico Identification and

Mechanism Exploration of Hepatotoxic Ingredients in Traditional Chinese Medicine. *Front. Pharmacol.* **2019**, *10* (May), 1–15. https://doi.org/10.3389/fphar.2019.00458.

(72) Takeshita, J. ichi; Nakayama, H.; Kitsunai, Y.; Tanabe, M.; Oki, H.; Sasaki, T.; Yoshinari, K. Discriminative Models Using Molecular Descriptors for Predicting Increased Serum ALT Levels in Repeated-Dose Toxicity Studies of Rats. *Comput. Toxicol.* **2018**, *6*, 64–70. https://doi.org/10.1016/j.comtox.2017.05.002.

(73) Toropova, A. P.; Toropov, A. A. CORAL: Binary Classifications (Active/Inactive) for Drug-Induced Liver Injury. *Toxicol. Lett.* **2017**, *268*, 51–57. https://doi.org/10.1016/j.toxlet.2017.01.011.

(74) Zhang, H.; Ding, L.; Zou, Y.; Hu, S. Q.; Huang, H. G.; Kong, W. B.; Zhang, J. Predicting Drug-Induced Liver Injury in Human with Naïve Bayes Classifier Approach. *J. Comput. Aided. Mol. Des.* **2016**, *30* (10), 889–898. https://doi.org/10.1007/s10822-016-9972-6.

(75) Schöning, V.; Krähenbühl, S.; Drewe, J. The Hepatotoxic Potential of Protein Kinase Inhibitors Predicted with Random Forest and Artificial Neural Networks. *Toxicol. Lett.* **2018**, *299* (October), 145–148. https://doi.org/10.1016/j.toxlet.2018.10.009.

(76) Chen, M.; Borlak, J.; Tong, W. A Model to Predict Severity of Drug-Induced Liver Injury in Humans. *Hepatology* **2016**, *64* (3), 931–940. https://doi.org/10.1002/hep.28678.

(77) Hammann, F.; Schöning, V.; Drewe, J. Prediction of Clinically Relevant Drug-Induced Liver Injury from Structure Using Machine Learning. *J. Appl. Toxicol.* **2019**, *39* (3), 412–419. https://doi.org/10.1002/jat.3741.

(78) Chen, M.; Borlak, J.; Tong, W. High Lipophilicity and High Daily Dose of Oral Medications Are Associated with Significant Risk for Drug-Induced Liver Injury. *Hepatology* **2013**, *58* (1), 388–396. https://doi.org/10.1002/hep.26208.

(79) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.

(80) Galar, M.; Fern, A.; Barrenechea, E.; Bustince, H. A Review of Ensembles for the Class Imbalance Problem. **2012**, *42* (4), 463–484.

(81) Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling Imbalanced Datasets : A Review. *Science (80-.).* **2006**, *30* (1), 25–36. https://doi.org/10.1007/978-0-387-09823-4_45.

(82) Kotsampasakou, E.; Brenner, S.; Jäger, W.; Ecker, G. F. Identification of Novel Inhibitors of Organic Anion Transporting Polypeptides 1B1 and 1B3 (OATP1B1 and OATP1B3) Using a Consensus Vote of Six Classification Models. *Mol. Pharm.* **2015**, *12* (12), 4395–4404. https://doi.org/10.1021/acs.molpharmaceut.5b00583.

(83) Wang, H.; Liu, R.; Schyman, P.; Wallqvist, A. Deep Neural Network Models for Predicting Chemically Induced Liver Toxicity Endpoints from Transcriptomic Responses. *Front. Pharmacol.* **2019**, *10* (FEB), 1–12. https://doi.org/10.3389/fphar.2019.00042.

(84) Kotsampasakou, E.; Ecker, G. F. Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters—An in Silico Modeling Approach. *J. Chem. Inf. Model.* **2017**, *57* (3), 608–615. https://doi.org/10.1021/acs.jcim.6b00518.

(85) Jain, S.; Kotsampasakou, E.; Ecker, G. F. Comparing the Performance of Meta-Classifiers—a Case Study on Selected Imbalanced Data Sets Relevant for Prediction of Liver Toxicity. *J. Comput. Aided. Mol. Des.* **2018**, *32* (5), 583–590. https://doi.org/10.1007/s10822-018-0116-z.

(86) Breiman, L. Bagging Predictions. *Mach. Learn.* **1996**, *24* (2), 123–140.

(87) Domingos, P. MetaCost: A General Method for Making Classifiers Cost-Sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*; ACM Press: New York, New York, USA, 1999; Vol. 25, pp 155–164. https://doi.org/10.1145/312129.312220.

(88) Egan, W. J.; Zlokarnik, G.; Grootenhuis, P. D. J. In Silico Prediction of Drug Safety: Despite Progress There Is Abundant Room for Improvement. *Drug Discov. Today Technol.* **2004**, *1* (4), 381–387. https://doi.org/10.1016/j.ddtec.2004.11.002.

(89) Hewitt, M.; Enoch, S. J.; Madden, J. C.; Przybylak, K. R.; Cronin, M. T. D. Hepatotoxicity: A Scheme for Generating Chemical Categories for Read-across, Structural Alerts and Insights into Mechanism(s) of Action. *Crit. Rev. Toxicol.* **2013**, *43* (7), 537–558. https://doi.org/10.3109/10408444.2013.811215.

(90) Pizzo, F.; Lombardo, A.; Manganaro, A.; Benfenati, E. A New Structure-Activity Relationship (SAR) Model for Predicting Drug-Induced Liver Injury, Based on Statistical and Expert-Based Structural Alerts. *Front. Pharmacol.* **2016**, *7* (NOV), 1–15. https://doi.org/10.3389/fphar.2016.00442.

(91) Liu, R.; Yu, X.; Wallqvist, A. Data-Driven Identification of Structural Alerts for Mitigating the Risk of Drug-Induced Human Liver Injuries. *J. Cheminform.* **2015**, *7* (1), 4. https://doi.org/10.1186/s13321-015-0053-y.

(92) Clark, R. D.; Wolohan, P. R. N.; Hodgkin, E. E.; Kelly, J. H.; Sussman, N. L. Modelling in Vitro Hepatotoxicity Using Molecular Interaction Fields and SIMCA. *J. Mol. Graph. Model.* **2004**, *22* (6), 487–497. https://doi.org/10.1016/j.jmgm.2004.03.009.

(93) Matthews, E. J.; Ursem, C. J.; Kruhlak, N. L.; Benz, R. D.; Sabaté, D. A.; Yang, C.; Klopman, G.; Contrera, J. F. Identification of Structure-Activity Relationships for Adverse Effects of Pharmaceuticals in Humans: Part B. Use of (Q)SAR Systems for Early Detection of Drug-Induced Hepatobiliary and Urinary Tract Toxicities. *Regul. Toxicol. Pharmacol.* **2009**, *54* (1), 23–42. https://doi.org/10.1016/j.yrtph.2009.01.009.

(94) Liu, Z.; Shi, Q.; Ding, D.; Kelly, R.; Fang, H.; Tong, W. Translating Clinical Findings into Knowledge in Drug Safety Evaluation--Drug Induced Liver Injury Prediction System (DILIps). *PLoS Comput. Biol.* **2011**, *7* (12), e1002310. https://doi.org/10.1371/journal.pcbi.1002310.

(95) Myshkin, E.; Brennan, R.; Khasanova, T.; Sitnik, T.; Serebriyskaya, T.; Litvinova, E.; Guryanov, A.; Nikolsky, Y.; Nikolskaya, T.; Bureeva, S. Prediction of Organ Toxicity Endpoints by QSAR Modeling Based on Precise Chemical-Histopathology Annotations. *Chem. Biol. Drug Des.* **2012**, *80* (3), 406–416. https://doi.org/10.1111/j.1747-0285.2012.01411.x.

(96) Chen, M.; Hong, H.; Fang, H.; Kelly, R.; Zhou, G.; Borlak, J.;

Tong, W. Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation and Using a Large Collection of Drugs. *Toxicol. Sci.* **2013**, *136* (1), 242–249. https://doi.org/10.1093/toxsci/kft189.

(97) Zhu, X. W.; Sedykh, A.; Liu, S. S. Hybrid in Silico Models for Drug-Induced Liver Injury Using Chemical Descriptors and in Vitro Cell-Imaging Information. *J. Appl. Toxicol.* **2014**, *34* (3), 281–288. https://doi.org/10.1002/jat.2879.

(98) Zhu, X. W.; Xin, Y. J.; Chen, Q. H. Chemical and in Vitro Biological Information to Predict Mouse Liver Toxicity Using Recursive Random Forests. *SAR QSAR Environ. Res.* **2016**, *27* (7), 559–572. https://doi.org/10.1080/1062936X.2016.1201142.

(99) Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Prediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. *Mol. Inform.* **2016**, *35* (3–4), 136–144. https://doi.org/10.1002/minf.201500055.

(100) Hong, H.; Thakkar, S.; Chen, M.; Tong, W. Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-Approved Drugs. *Sci. Rep.* **2017**, *7* (1), 17311. https://doi.org/10.1038/s41598-017-17701-7.

(101) Kim, E.; Nam, H. Prediction Models for Drug-Induced Hepatotoxicity by Using Weighted Molecular Fingerprints. *BMC Bioinformatics* **2017**, *18* (Suppl 7). https://doi.org/10.1186/s12859-017-1638-4.

(102) McEuen, K.; Borlak, J.; Tong, W.; Chen, M. Associations of Drug Lipophilicity and Extent of Metabolism with Drug-Induced Liver Injury. *Int. J. Mol. Sci.* **2017**, *18* (7). https://doi.org/10.3390/ijms18071335.

(103) Wu, L.; Liu, Z.; Auerbach, S.; Huang, R.; Chen, M.; McEuen, K.; Xu, J.; Fang, H.; Tong, W. Integrating Drug's Mode of Action into Quantitative Structure-Activity Relationships for Improved Prediction of Drug-Induced Liver Injury. *J. Chem. Inf. Model.* **2017**, *57* (4), 1000–1006. https://doi.org/10.1021/acs.jcim.6b00719.

(104) Li, X.; Chen, Y.; Song, X.; Zhang, Y.; Li, H.; Zhao, Y. The Development and Application of in Silico Models for Drug Induced Liver Injury. *RSC Adv.* **2018**, *8* (15), 8101–8111. https://doi.org/10.1039/c7ra12957b.

(105) Papa, E.; Sangion, A.; Taboureau, O.; Gramatica, P. Quantitative Prediction of Rat Hepatotoxicity by Molecular Structure. *Int. J. Quant. Struct. Relationships* **2018**, *3* (2), 49–60. https://doi.org/10.4018/ijqspr.2018070104.

(106) Ambe, K.; Ishihara, K.; Ochibe, T.; Ohya, K.; Tamura, S.; Inoue, K.; Yoshida, M.; Tohkin, M. In Silico Prediction of Chemical-Induced Hepatocellular Hypertrophy Using Molecular Descriptors. *Toxicol. Sci.* **2018**, *162* (2), 667–675. https://doi.org/10.1093/toxsci/kfx287.

(107) López-Massaguer, O.; Pinto-Gil, K.; Sanz, F.; Amberg, A.; Anger, L. T.; Stolte, M.; Ravagli, C.; Marc, P.; Pastor, M. Generating Modeling Data from Repeat-Dose Toxicity Reports. *Toxicol. Sci.* **2018**, *162* (1), 287–300. https://doi.org/10.1093/toxsci/kfx254.

(108) Chan, R.; Benet, L. Z. Evaluation of DILI Predictive Hypotheses in Early Drug Development. *Chem. Res. Toxicol.* **2017**, *30* (4), 1017–1029. https://doi.org/10.1021/acs.chemrestox.7b00025.

(109) He, S.; Ye, T.; Wang, R.; Zhang, C.; Zhang, X.; Sun, G.; Sun, X. An In Silico Model for Predicting Drug-Induced Hepatotoxicity. *Int. J. Mol. Sci.* **2019**, *20* (8), 1897. https://doi.org/10.3390/ijms20081897.

(110) Wang, Y.; Xiao, Q.; Chen, P.; Wang, B. In Silico Prediction of Drug-Induced Liver Injury Based on Ensemble Classifier Method. *Int. J. Mol. Sci.* **2019**, *20* (17), 4106. https://doi.org/10.3390/ijms20174106.

(111) Aleo, M. D.; Shah, F.; Allen, S.; Barton, H. A.; Costales, C.; Lazzaro, S.; Leung, L.; Nilson, A.; Obach, R. S.; Rodrigues, A. D.; et al. Moving Beyond Binary Predictions of Human Drug-Induced Liver Injury (Dili) Towards Contrasting Relative Risk Potential. *Chem. Res. Toxicol.* **2019**, No. Dili. https://doi.org/10.1021/acs.chemrestox.9b00262.

(112) Williams, D. P.; Lazic, S.; Foster, A. J.; Semenova, E.; Morgan, P. Predicting Drug-Induced Liver Injury with Bayesian Machine Learning. *Chem. Res. Toxicol.* **2019**, acs.chemrestox.9b00264. https://doi.org/10.1021/acs.chemrestox.9b00264.

(113) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. **2016**. https://doi.org/10.1145/2939672.2939785.

(114) Klekota, J.; Roth, F. P. Chemical Substructures That Enrich for Biological Activity. *Bioinformatics* **2008**, *24* (21), 2518–2525. https://doi.org/10.1093/bioinformatics/btn479.

(115) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J Chem Inf Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(116) Cronin, M. T. D.; Enoch, S. J.; Mellor, C. L.; Przybylak, K. R.; Richarz, A. N.; Madden, J. C. In Silico Prediction of Organ Level Toxicity: Linking Chemistry to Adverse Effects. *Toxicol. Res.* **2017**, *33* (3), 173–182. https://doi.org/10.5487/TR.2017.33.3.173.

(117) Ellison, C. M.; Madden, J. C.; Judson, P.; Cronin, M. T. D. Using in Silico Tools in a Weight of Evidence Approach to Aid Toxicological Assessment. *Mol. Inform.* **2010**, *29* (1–2), 97–110. https://doi.org/10.1002/minf.200900006.

(118) Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D. Structural Alert/Reactive Metabolite Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24* (9), 1345–1410. https://doi.org/10.1021/tx200168d.

(119) Yu, K.; Geng, X.; Chen, M.; Zhang, J.; Wang, B.; Ilic, K.; Tong, W. High Daily Dose and Being a Substrate of Cytochrome P450 Enzymes Are Two Important Predictors of Drug-Induced Liver Injury. *Drug Metab. Dispos.* **2014**, *42* (4), 744–750. https://doi.org/10.1124/dmd.113.056267.

(120) Leeson, P. D.; St-Gallay, S. A.; Wenlock, M. C. Impact of Ion Class and Time on Oral Drug Molecular Properties. *Medchemcomm* **2011**, *2* (2), 91–105. https://doi.org/10.1039/c0md00157k.

(121) García-Cortés, M.; Lucena, M. I.; Pachkoria, K.; Borraz, Y.; Hidalgo, R.; Andrade, R. J. Evaluation of Naranjo Adverse Drug Reactions Probability Scale in Causality Assessment of Drug-Induced Liver Injury. *Aliment. Pharmacol. Ther.* **2008**, *27* (9), 780–789. https://doi.org/10.1111/j.1365-

2036.2008.03655.x.

(122) Björnsson, E. S.; Hoofnagle, J. H. Categorization of Drugs Implicated in Causing Liver Injury: Critical Assessment Based on Published Case Reports. *Hepatology* **2016**, *63* (2), 590–603. https://doi.org/10.1002/hep.28323.

(123) Teschke, R.; Eickhoff, A.; Frenzel, C.; Wolff, A.; J., S. Drug Induced Liver Injury: Accuracy of Diagnosis in Published Reports. *Ann. Hepatol.* **2014**, *13* (2), 248–255.

(124) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**.

(125) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom-Centred Fragments. *J. Chem. Soc. C Org.* **1971**, No. 3702, 3702. https://doi.org/10.1039/j39710003702.

(126) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. *Match* **2006**, *56* (2), 237–248.

(127) Tetko, I. V; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity Against. *Osiris* **2008**, 1733–1746.

(128) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; et al. DrugBank 3.0: A Comprehensive Resource for "omics" Research on Drugs. *Nucleic Acids Res.* **2011**, *39* (Database issue), D1035-41. https://doi.org/10.1093/nar/gkq1126.

(129) Jain, S.; Ecker, G. F. In Silico Approaches to Predict Drug-Transporter Interaction Profiles: Data Mining, Model Generation, and Link to Cholestasis. In *Experimental Cholestasis Research*; Vinken, M., Ed.; Humana Press: New York, 2019; pp 383–396. https://doi.org/10.1007/978-1-4939-9420-5_26.

(130) Welch, M. A.; Kock, K.; Urban, T. J.; Brouwer, K. L. R.; Swaan, P. W. Toward Predicting Drug-Induced Liver Injury: Parallel Computational Approaches to Identify Multidrug Resistance Protein 4 and Bile Salt Export Pump Inhibitors. *Drug Metab. Dispos.* **2015**, *43* (5), 725–734. https://doi.org/10.1124/dmd.114.062539.

(131) Montanari, F.; Pinto, M.; Khunweeraphong, N.; Wlcek, K.; Sohail, M. I.; Noeske, T.; Boyer, S.; Chiba, P.; Stieger, B.; Kuchler, K.; et al. Flagging Drugs That Inhibit the Bile Salt Export Pump. *Mol. Pharm.* **2016**, *13* (1), 163–171. https://doi.org/10.1021/acs.molpharmaceut.5b00594.

(132) Montanari, F.; Zdrazil, B.; Digles, D.; Ecker, G. F. Selectivity Profiling of BCRP versus P-Gp Inhibition: From Automated Collection of Polypharmacology Data to Multi-Label Learning. *J. Cheminform.* **2016**, *8* (1), 1–13. https://doi.org/10.1186/s13321-016-0121-y.

(133) Dawson, S.; Stahl, S.; Paul, N.; Barber, J.; Kenna, J. G. In Vitro Inhibition of the Bile Salt Export Pump Correlates with Risk of Cholestatic Drug-Induced Liver Injury in Humans. *Drug Metab. Dispos.* **2012**, *40* (1), 130–138. https://doi.org/10.1124/dmd.111.040758.

(134) Xi, L.; Yao, J.; Wei, Y.; Wu, X.; Yao, X.; Liu, H.; Li, S. The in Silico Identification of Human Bile Salt Export Pump (ABCB11) Inhibitors Associated with Cholestatic Drug-Induced Liver Injury. *Mol. Biosyst.* **2017**, *13* (2), 417–424. https://doi.org/10.1039/c6mb00744a.

(135) Warner, D. J.; Chen, H.; Cantin, L.-D.; Kenna, J. G.; Stahl, S.; Walker, C. L.; Noeske, T. Mitigating the Inhibition of Human Bile Salt Export Pump by Drugs: Opportunities Provided by Physicochemical Property Modulation, In Silico Modeling, and Structural Modification. *Drug Metab. Dispos.* **2012**, *40* (12), 2332–2341. https://doi.org/10.1124/dmd.112.047068.

(136) Pauli-Magnus, C.; Meier, P. J. Hepatocellular Transporters and Cholestasis. *J. Clin. Gastroenterol.* **2005**, *39* (4 SUPPL.), 103–110.

(137) Gadaleta, D.; Manganelli, S.; Roncaglioni, A.; Toma, C.; Benfenati, E.; Mombelli, E. QSAR Modeling of ToxCast Assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. *J. Chem. Inf. Model.* **2018**, *58* (8), 1501–1517. https://doi.org/10.1021/acs.jcim.8b00297.

(138) Mellor, C. L.; Marchese Robinson, R. L.; Benigni, R.; Ebbrell, D.; Enoch, S. J.; Firman, J. W.; Madden, J. C.; Pawar, G.; Yang, C.; Cronin, M. T. D. Molecular Fingerprint-Derived Similarity Measures for Toxicological Read-across: Recommendations for Optimal Use. *Regul. Toxicol. Pharmacol.* **2019**, *101* (October 2018), 121–134. https://doi.org/10.1016/j.yrtph.2018.11.002.

(139) Liu, Y. Incorporation of Absorption and Metabolism into Liver Toxicity Prediction for Phytochemicals: A Tiered in Silico QSAR Approach. *Food Chem. Toxicol.* **2018**, *118* (April), 409–415. https://doi.org/10.1016/j.fct.2018.05.039.

(140) Peng, Y.; Wu, Z.; Yang, H.; Cai, Y.; Liu, G.; Li, W.; Tang, Y. Insights into Mechanisms and Severity of Drug-Induced Liver Injury via Computational Systems Toxicology Approach. *Toxicol. Lett.* **2019**, *312* (April), 22–33. https://doi.org/10.1016/j.toxlet.2019.05.005.

(141) Su, R.; Wu, H.; Xu, B.; Liu, X.; Wei, L. Developing a Multi-Dose Computational Model for Drug-Induced Hepatotoxicity Prediction Based on Toxicogenomics Data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2018**, *5963* (c). https://doi.org/10.1109/TCBB.2018.2858756.

(142) Wu, Y.; Wang, G. Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. *Int. J. Mol. Sci.* **2018**, *19* (8), 2358. https://doi.org/10.3390/ijms19082358.

(143) Battista, C.; Howell, B. A.; Siler, S. Q.; Watkins, P. B. An Introduction to DILIsym® Software, a Mechanistic Mathematical Representation of Drug-Induced Liver Injury; 2018; pp 101–121. https://doi.org/10.1007/978-1-4939-7677-5_6.

(144) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010. https://doi.org/10.1021/jm4004285.

(145) Wang, D. Infer the in Vivo Point of Departure with ToxCast in Vitro Assay Data Using a Robust Learning Approach. *Arch. Toxicol.* **2018**.

(146) Olson, H.; Betton, G.; Robinson, D.; Thomas, K.; Monro, A.; Kolaja, G.; Lilly, P.; Sanders, J.; Sipes, G.; Bracken, W.; et al. Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals. *Regul. Toxicol. Pharmacol.* **2000**, *32* (1), 56–67. https://doi.org/10.1006/rtph.2000.1399.

1
2