MASTERARBEIT | MASTER'S THESIS

Titel | Title

Application of Substructural and Biological Fingerprints for modeling Drug-induced Fatty Liver Disease

verfasst von | submitted by

Florian Oehler BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Magister pharmaciae (Mag. pharm.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 605
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Pharmazie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Gerhard Ecker, for giving me the opportunity to carry out this thesis in his research group. His constructive feedback and valuable discussions greatly contributed to the progress of this work.

I would also like to thank all members of the Pharmacoinformatics Research Group and fellow students who provided support and a pleasant working environment. My special thanks go to Mag. Palle Steen Helmke, whose continuous assistance and insightful discussions helped me during this thesis.

Finally, I am deeply grateful to my parents for their constant and unconditional support. I also wish to sincerely thank my family, friends and Franziska, who encouraged me throughout my entire studies.

Abstract

Objective

This thesis aimed to use a publicly available KNIME workflow to develop binary classification models for predicting drug-induced fatty liver disease (DIFLD) employing structural and readily interpretable biological fingerprints. The work further included the feature importance to enable potential hypotheses generation, and the evaluation of seven compounds of interest with the best-performing models.

Methods

Four binary classification datasets were employed: three for DIFLD (LJMU, FAERS+DILIst, and FAERS+FDA) and one for general drug-induced liver injury (DILI) (DILIst). While the LJMU dataset could only be encoded using structural information, the other combined PubChem Bitvectors (BV) with biological fingerprints (targets and pathways). Class imbalances were addressed utilizing Synthetic Minority Oversampling (SMOTE) and Undersampling (US). For machine learning (ML), Random Forest (RF) and gradient-boosted tree (GB) models were used. Visualization of feature and chemical space was done via t-SNE and MolCompass. Uncertainty filtering and distance-to-nearest-neighbor (DtNN) analysis were applied to assess confidence and applicability domain.

Results

The best-performing model based on the FAERS+DILIst dataset was only target-driven and achieved a sensitivity of 0.79 and a Matthews Correlation Coefficient (MCC) of 0.42. For the LJMU dataset, the best structure-based model reached a sensitivity of 0.47 and an MCC of 0.24. Nuclear receptors, Albumin, Prelamin A/C, CYP-enzymes, and ABC transporters were among the 30 most important features. Two of the best-performing models – one LJMU dataset model employing only substructural information and a FAERS+DILIst model which used all three fingerprints – were chosen to see how they would classify compounds of interest. For three molecules predictions were contradictory, with the LJMU model exhibiting notably lower uncertainties and a spatial proximity to a distinct cluster of DIFLD-positives.

Conclusion

Although the use of biological fingerprints and the designated workflow had limitations, models based on drug datasets still reached reasonable performance in ten-fold cross-

validation. The feature importance, particularly of the target fingerprint, provided interpretable results often consistent with literature findings.

Zusammenfassung

Zielsetzung

Ziel dieser Arbeit war es, einen öffentlich verfügbaren KNIME-Workflow zu nutzen, um binäre Klassifikationsmodelle zur Vorhersage von arzneimittelinduzierter Fettlebererkrankung (DIFLD) unter Verwendung struktureller und leicht interpretierbarer biologischer Fingerprints zu entwickeln. Darüber hinaus wurde die Feature Importance präsentiert, um die Generierung potenzieller Hypothesen zu ermöglichen, und sieben Verbindungen mit ausgewählten, leistungsstarken Modellen evaluiert.

Methoden

Es wurden vier binäre Klassifikationsdatensätze eingesetzt: drei für DIFLD (LJMU, FAERS+DILIst, and FAERS+FDA) und einer für arzneimittelinduzierte Leberschädigung (DILI) im Allgemeinen (DILIst). Während der LJMU-Datensatz ausschließlich mit Struktur-Information codiert werden konnte, kombinierten die anderen PubChem Bitvektoren (BV) mit biologischen Fingerprints (Targets und Signalwege). Unterschiede in der Anzahl an positiven und negativen Verbindungen innerhalb der Datensätze wurden mittels Synthetic Minority Oversampling (SMOTE) und Undersampling (US) ausgeglichen. Für das maschinelle Lernen (ML) kamen Random Forest (RF) und Gradient-Boosted Tree (GB) Modelle zum Einsatz. Die Visualisierung des Feature- und chemischen Raums erfolgte über t-SNE und MolCompass. Unsicherheitsfilterung und Distance-to-nearest-Neighbor (DtNN) Analyse wurden zur Bewertung der Klassifikationssicherheit und Anwendbarkeit herangezogen.

Ergebnisse

Das leistungsstärkste Modell, basierend auf dem FAERS+DILIst-Datensatz, war ausschließlich Target-basiert und erreichte eine Sensitivität von 0,79 sowie einen Matthews Korrelations-Koeffizient (MCC) von 0,42. Für den LJMU-Datensatz erreichte das beste struktur-basierte Modell eine Sensitivität von 0,47 und einen MCC von 0,24. Unter den 30 wichtigsten Features waren nukleäre Rezeptoren, Albumin, Prelamin A/C, CYP-Enzyme und ABC-Transporter. Zwei der besten Modelle – ein LJMU-Datensatzmodell, das nur auf struktureller Information basierte und ein FAERS+DILIst-Modell, das alle drei Fingerprinttypen vereinte – wurden ausgewählt, um zu sehen, wie sie die sieben ausgewählten Verbindungen klassifizieren würden. Für drei Moleküle ergaben sich widersprüchliche

Vorhersagen; jedoch zeigte das LJMU-Modell deutlich geringere Unsicherheiten und eine räumliche Nähe zu einem abgegrenzten Cluster aus DIFLD-Positiven.

Schlussfolgerung

Obwohl die Verwendung biologischer Fingerprints und des vorgesehenen Workflows Einschränkungen aufwies, konnten Modelle auf Basis von Arzneistoff-Datensätzen eine angemessene Performance in der Ten-Fold-Cross-Validation erreichen. Die Feature Importance – insbesondere bezogen auf den Target-Fingerprint – lieferte interpretierbare Ergebnisse, die teilweise mit der Literatur übereinstimmten.

Table of Contents

Abstract **3**
Zusammenfassung **5**
List of abbreviations **9**
List of figures **10**
List of tables **11**
1 Introduction **12**
1.1 Drug-induced liver injury and hepatic steatosis 12
1.2 Pathological mechanisms involved in drug-induced fatty liver disease 13
1.2 Computational methods for hepatotoxicity-prediction 15
1.3 Aim of this thesis 17
2 Methods **18**
2.1 KNIME 18
2.2 Binary datasets 18
2.2.1 LJMU 18
2.2.2 FAERS+DILIst 19
2.2.3 FAERS+FDA 20
2.2.4 DILIst 20
2.3 Target fingerprint 21
2.4 Tissue filtering 22
2.5 Pathway fingerprint 23
2.6 PubChem bitvectors for structural fingerprint 24
2.7 Machine Learning 25
2.8 Dealing with class imbalances 26
2.9 Performance evaluation 27
2.10 Feature space visualization via traditional t-SNE 28
2.11 Chemical space visualization via MolCompass 29
2.12 Uncertainty and applicability domain 29
3 Results **32**
3.1 Fingerprint length 32
3.2 Comparison of the FAERS-derived DIFLD-positives with DILI-positives 33
3.2 Model performances 34

3.3 Number and Distribution of active Fingerprints per compound 39
3.4 Feature Importance and distribution of actives 41
3.5 A visualization of the feature space 48
3.6 A visualization of the chemical space.....50
3.8 Prediction Results, Uncertainty and Applicability 52
3.9 Uncertainty filtering 55
4 Discussion 57
4.1 Datasets.....57
4.2 Model Performances.....58
4.3 Feature importance.....60
4.4 A cluster of DIFLD-positives in the feature space 62
4.4 Limitations of this thesis.....63
5 Conclusion and outlook.....65
References 67

List of abbreviations

ABC ..... ATP-binding cassette
AhR ..... aryl hydrocarbon receptor
AOP ..... adverse outcome pathway
BA ..... balanced accuracy
BCRP ..... breast cancer resistance protein
BSEP ..... bile salt export pump
BV ..... bitvector
CAR ..... constitutive androstane receptor
CPT-1 ..... Carnitine palmitoyl-transferase 1
DIFLD ..... drug-induced fatty liver disease
DILI ..... drug-induced liver injury
DILIN ..... drug-induced liver injury network
DISH ..... drug-induced steatohepatitis
DtNN ..... distance-to-nearest-neighbor
ECFP ..... extended-connectivity fingerprint
FAERS ..... FDA adverse event reporting system
FDR ..... false-discovery rate
FXR ..... farnesoid x receptor
GB ..... gradient boost
GR ..... glucocorticoide receptor
Imb ..... imbalanced
LTKB ..... liver toxicity knowledge base
LXR ..... liver x receptor
MASH ..... metabolic dysfunction-associated steatohepatitis
MASLD ..... metabolic dysfunction-associated steatotic liver disease
MCC ..... matthews correlation coefficient
MIE ..... molecular initiating event
ML ..... machine learning
MRP ..... multidrug resistance-related protein
mtFAO ..... mitochondrial fatty acid oxidation
MTP ..... microsomal triglyceride transfer protein
NAFLD ..... non-alcoholic fatty liver disease
NAM ..... new approach methodologies
NASH ..... non-alcoholic steatohepatitis
NRF2 ..... nuclear factor erythroid 2-related factor 2
nTPM ..... normalized transcripts per million
OATP ..... organic anione transporter polypeptide
P-gp ..... permeability glycoprotein
PPAR ..... peroxisome proliferator-activated receptor gamma
PXR ..... pregnane x receptor
QSAR ..... quantitative structure-activity relationships
RF ..... random forest
SAR ..... structure-activity relationship
SD ..... standard deviation
SMOTE ..... synthetic minority oversampling technique
t-SNE ..... t-distributed stochastic neighbor embedding
UC ..... uncertainty
UGT ..... uridine 5'-diphospho-glucuronosyltransferase
US ..... undersampling

List of figures

Figure 1: Excerpt of the generated PubChem target data retrieval workflow. 22
Figure 2: Schematic overview of the designated KNIME workflow. 25
Figure 3: Example table view in KNIME illustrating the combined fingerprint matrix.. 25
Figure 4: Schematic excerpt of the generated feature visualization workflow.. 29
Figure 5: Excerpt of the generated applicability domain workflow.. 31
Figure 6: Table view in KNIME illustrating the Tanimoto distances 31
Figure 7: Bar chart illustrating the comparison of DIFLD-positives with DILI-positives. 33
Figure 8: Boxplot illustrating the distribution of the active target-count per compound 40
Figure 9: Screenshot of the interactive feature-space scatter plot of LJMU 49
Figure 10: Screenshot of the interactive feature-space scatter plot of FAERS+DILIst 50
Figure 11: Comparison of the chemical-space scatter plots 51
Figure 12: Feature-space scatter plot of LJMU compounds (PubChem BV).. 54
Figure 13: Zoomed-in excerpt of the feature-space scatter plot of LJMU 62

List of tables

Table 1. Sizes and class balances of datasets. 32
Table 2. Fingerprint lengths of datasets 33
Table 3. Ten-fold cross-validation results of the LJMU dataset models 34
Table 4. Ten-fold cross-validation results of the FAERS+DILIst dataset models 35
Table 5. Ten-fold cross-validation results of the FAERS+FDA dataset models 36
Table 6. Ten-fold cross-validation results of the DILIst dataset models 38
Table 7. Distribution of active target-count per compound 39
Table 8. Impact of active-target count-based compound filtering 41
Table 9. Top 5 most important features of LJMU SMOTE_GB model 42
Table 10. Top 30 most important features of FAERS+DILIst SMOTE_RF model 43
Table 11. Top 30 most important features of FAERS+DILIst US_GB model 44
Table 12. Top 30 most important features of DILIst US_GB model 46
Table 13. Top 30 most important features of DILIst SMOTE_RF model 47
Table 14. Prediction outcomes, uncertainty and applicability domain 1 53
Table 15. Prediction outcomes, uncertainty and applicability domain 2 55
Table 16. Impact of uncertainty filtering on cross-validation results of LJMU 56
Table 17. Impact of uncertainty filtering on cross-validation results of FAERS+DILIst 56

1 Introduction

1.1 Drug-induced liver injury and hepatic steatosis

Drug-Induced Liver Injury (DILI) is a broadly applicable term that can be described as liver damage due to exposure to potentially hepatotoxic compounds (Allison et al., 2023). DILI is a substantial issue in early- and preclinical drug development, being a primary cause for attrition of drug candidates (Korver et al., 2021). Moreover, one third of the post-marketing withdrawals of drugs because of adverse reactions are on account of DILI (Dirven et al., 2021). Although, it is a rare phenomenon with an incidence of 2.3-23.8 cases per 100.00 persons per year (Li et al., 2022), DILI is the leading cause of acute liver failure in the western world, associated with a case fatality ranging from 10-50% (Hosack et al., 2023). A distinction is being made between intrinsic (predictable) DILI and idiosyncratic (unpredictable) DILI, which is the more frequently occurring one. While intrinsic DILI is dose-dependent and reproducible in animal models, idiosyncratic DILI is non-dose-dependent and in most cases not reproducible in any experimental animal models (Allison et al., 2023; Katarey & Verma, 2016), highlighting the need for alternative approaches.

Hepatic Steatosis, which is also known as fatty liver disease, is defined by the hepatic accumulation of lipids. Since minor fat deposition in the liver also occurs in healthy individuals, at least 5% of the hepatocytes must be visually affected for the condition to be considered pathological (Borgne-Sanchez & Fromenty, 2025; Sharma & John, 2025). The increased fat deposition is due to disruptions in lipid synthesis, transport, clearance or a combination of these processes. Hepatic steatosis presents in two forms: macrovesicular steatosis, where large lipid droplets displace the hepatocyte nucleus to the cell periphery, and microvesicular steatosis, where smaller droplets occur without nucleus displacement (Ortega-Vallbona et al., 2024).

Metabolic dysfunction-associated steatotic liver disease (MASLD, previously referred to as non-alcoholic fatty liver disease NAFLD) describes a spectrum of liver abnormalities, from simple hepatic fat accumulation (steatosis) to more advanced stages including inflammation and hepatocellular injury (steatohepatitis). MASLD is associated with metabolic conditions such as obesity, insulin resistance, hyperglycemia, dyslipidemia, hypertension and cardiometabolic risk factors (López-Pascual et al., 2024). Even though it has been estimated that fewer than 2% of metabolic dysfunction-associated steatohepatitis

(MASH, formerly non-alcoholic steatohepatitis NASH) cases are caused by drugs (García-Cortés et al., 2023), findings indicate that drugs can mimic and exacerbate MASLD. Furthermore, certain drugs that are frequently used in clinical practice (such as ibuprofen, naproxen, acetaminophen, irinotecan, methotrexate, and tamoxifen) are expected to promote/worsen hepatic steatosis in individuals with already existing MASLD. As stated by López-Pascual et al., a bi-directional link between DILI and MASLD seems likely: while some drugs act as pro-steatogenic agents, pre-existing MASLD may make hepatocytes more susceptible to drug-induced damage. Thus, patients with underlying MASLD are at elevated risk of developing DILI when exposed to these drugs (López-Pascual et al., 2024).

As reported by the Drug-Induced Liver Injury Network (DILIN), around 27% of DILI cases include some variation of steatosis (Kolaric et al., 2021). However, since pre-existing MASLD / hepatic steatosis may increase hepatocellular susceptibility to drug-induced injury, it could be unclear whether the steatosis was present before drug exposure or developed because of it (López-Pascual et al., 2024).

Drug-induced Fatty Liver Disease (DIFLD) and Drug-induced Steatohepatitis (DISH) are more specific terms which are also being used in this context (Kolaric et al., 2021; López-Pascual et al., 2024). For the sake of simplicity and readability, DIFLD is used in this thesis to overall refer to drug-induced hepatic steatosis and steatohepatitis.

1.2 Pathological mechanisms involved in drug-induced fatty liver disease

Since DILI is a more generic term which can stand for different specific endpoints such as vasculitic, neoplastic, cholestatic and steatohepatitic liver injury (Hosack et al., 2023), postulated mechanisms of DILI may vary a lot. Taking this into consideration, the application of more specific terms such as DIFLD seems to be important when talking about the potential mechanisms of drug-induced liver pathologies. Because this thesis focuses on drug-induced hepatic steatosis, DIFLD will be the main subject.

Mitochondrial dysfunction is thought to be a primary mechanism of DIFLD (Borgne-Sanchez & Fromenty, 2025). Particularly, drug-induced disruption of mitochondrial fatty acid oxidation (mtFAO) inhibits the oxidative breakdown of free fatty acids to acetyl-CoA. This leads to a hepatocellular accumulation of these fatty acids and ultimately to hepatic steatosis. Several drugs that are known to cause DIFLD (such as amiodarone, tamoxifen

and Irinotecan) can induce impairment of mtFAO (Kolaric et al., 2021). As stated by Borgne-Sanchez & Fromenty, microvesicular steatosis is associated with a strong inhibition of mtFAO, in contrast to macrovesicular steatosis which is linked to a milder dysfunction of beta-oxidation.

Other proposed mechanisms for DIFLD include an upregulated fatty acid influx, a promotion of de novo lipogenesis, an impairment of fatty acid transport across mitochondrial membranes and a reduced lipid export via VLDL (Kolaric et al., 2021; López-Pascual et al., 2024; Ortega-Vallbona et al., 2024). Carnitine palmitoyl-transferase 1 (CPT-1) is crucial for mitochondrial uptake of fatty acids from the cytosol so that the beta-oxidation can take place. While steatogenic compounds such as valproate, tamoxifen and amiodarone inhibit CPT-1, Ibuprofen and salicylic acid seem to trap CoA or L-carnitine, which are important cofactors for said uptake of fatty acids through mitochondrial membranes. Consequently, this may further contribute to the impairment of mtFAO (López-Pascual et al., 2024).

Since the microsomal triglyceride transfer protein (MTP) plays a role in the hepatic assembly of VLDL, the inhibition of MTP results in a reduced lipid export via VLDL. Thus, drugs which evidently inhibit MTP, such as tetracycline and tianeptine, could lead to an increased hepatocellular lipid storage (López-Pascual et al., 2024).

Besides these mechanisms, many nuclear receptors are thought to be involved in DIFLD. Several Adverse-Outcome-Pathway (AOP) networks contain these receptors as molecular initiating events (MIEs) which are started by interaction of the nuclear receptors with certain steatogenic chemicals (Ortega-Vallbona et al., 2024). The MIEs include the liver x receptor (LXR), farnesoid x receptor (FXR), constitutive androstane receptor (CAR), pregnane x receptor (PXR), estrogen receptor (ER), peroxisome proliferator-activated receptors (PPARs) and the glucocorticoid receptor (GR) (AbdulHameed et al., 2019; Ortega-Vallbona et al., 2024). These nuclear receptors are, for example, involved in events such as lipid entry/efflux, de novo lipogenesis, and also mtFAO (AbdulHameed et al., 2019). More specifically, PPARγ, LXR and PXR act as transcription factors, promoting de novo lipid synthesis and thus, upon activation, also potentially induce hepatic steatosis (López-Pascual et al., 2024).

AbdulHameed et al., who investigated public toxicogenomic data, stated that nuclear receptor activation (e.g. activation of FXR) might not be directly linked to steatosis and suggested that mitochondrial dysfunction itself may be a more relevant MIE. Moreover, some

reports seem to indicate that nuclear receptor interaction alone does not always lead to steatosis (AbdulHameed et al., 2019). However, it is important to remember that the MIEs represent not necessarily the activation of these nuclear receptors, but sometimes rather the inhibition, which can lead to wrong conclusions. One example of this is indeed FXR, where several studies have shown that FXR activation actually protects against hepatic steatosis (Clifford et al., 2021; Neuschwander-Tetri et al., 2015), while another study has reported that a reduced expression of FXR seems to worsen steatohepatitis (Armstrong & Guo, 2017).

1.2 Computational methods for hepatotoxicity-prediction

Both, in vitro assays and in vivo models are used to determine hepatotoxicity in preclinical testing. While in vitro experiments are based on cytochrome P450 (CYP450) inducibility, mitochondrial impairment or bile salt export pump (BSEP) inhibition, in vivo models rely on chimeric mice with humanized hepatocytes. However, these approaches sometimes fail when extrapolating to humans. Moreover, since the incidence of DILI is low and consequently large sample sizes are necessary, detection of DILI in clinical trials also remains a difficult endeavor (Lin et al., 2022). Idiosyncratic (unpredictable) DILI, as already mentioned, represents a particular challenge since it is non-dose-dependent and in most cases not reproducible in any experimental animal models (Katarey & Verma, 2016). So-called New Approach Methodologies (NAMs), including computational methods, such as machine learning (ML) systems, artificial intelligence (AI) or omics-based toxicity tests, provide an ethical, time-saving and cost-efficient alternative to conventional in vivo testing (Ortega-Vallbona et al., 2024). More specifically, used computational Methods to predict DILI or DIFLD include automatized structure-activity relationships (SARs), quantitative structure-activity relationship models (QSAR), read-across, omics, and structure-based approaches. SAR is based on molecular substructures which are thought to be responsible for certain biological or toxicological activities. Unlike SAR, QSAR models can also offer quantitative predictions such as affinity values or IC50 values (Ortega-Vallbona et al., 2024). While QSAR models may use structural alerts to encode the compounds, it can also be based on biological activity descriptors or pathway profiles (Füzi et al., 2023; Liu et al., 2021).

For predicting hepatotoxicity, Füzi et al. used binary fingerprints to encode DILI positive and negative compounds. These fingerprints included known target, interactome and pathway information from publicly available databases, such as ChEMBL and Reactome. Ultimately leading to binary matrices, where 0 indicates that there is no known connection between the compound and the biological entity and 1 stands for an existing interaction. Doing so, an accuracy of around 0.69 was achieved. Furthermore, the feature importance, which is basically a ranking of the fingerprints by their predicting significance, was analyzed. The top ranked fingerprints were then used to assess possible mechanisms for hepatotoxicity by comparing them to available evidence and literature. These fingerprints included, among others, fatty acid metabolism, arachidonic acid metabolism, heme degradation and CYP2C9 (Füzi et al., 2023).

Another study, which used structural alerts and predicted protein targets as descriptors for modeling DILI, also analyzed the feature importance. Since they likewise used the FDA DILIrank dataset (M. Chen et al., 2016) as a basis for their binary classification model, similar results were achieved. CYP2C9, CYP1A2, proteins involved in arachidonic acid metabolism and fatty acid metabolism were among the higher-ranking features. Moreover, their highest-ranked proteins included adenosine A1 receptor (GPCR), AR (androgen receptor) and oxidoreductases, such as aldose reductase (Liu et al., 2021).

Jain et al. created binary classification models specifically for predicting drug-induced hepatic steatosis using physicochemical descriptors (RDKit), ToxPrint features (chemical substructures), predictions from in-silico nuclear receptor models and in-silico transporter models. Nuclear receptor activation or inactivation was specifically predicted for PXR, aryl hydrocarbon receptor (AhR), LXR, PPARγ and the nuclear factor (erythroid-derived 2)-like 2 (NRF2). For transporters, the inhibition of P-gp, BCRP, BSEP, MRP3, MRP4 and OATPs was predicted. However, while the physicochemical and structural descriptors achieved reasonable results for predicting hepatic steatosis, the models based only on activity prediction data performed poorly (Jain et al., 2021).

Using the freely available Path4Drug workflow (Füzi et al., 2021), Helmke & Ecker applied interacting targets and pathways as biological fingerprints for a binary hepatotoxicity prediction model. More specifically, another subcategory of DILI, drug-induced cholestasis (DIC), was the endpoint of this model. In addition to targets and pathways, structure-based PubChem bitvectors (BV) and hepatic transporter inhibition predictions were used to encode the compounds. While all fingerprint-types led to reasonable DIC prediction

performances, a combination of substructure, pathway information and hepatic transporter inhibition predictions achieved the highest balanced accuracy (BA) (Helmke & Ecker, 2025).

1.3 Aim of this thesis

The aim of this thesis was to use and adapt the publicly available, cholestasis-based KNIME-workflow from Helmke & Ecker, which is based on the original work of Füzi et al., in an effort to create binary classification models which specifically predict DIFLD based on structural and readily interpretable biological fingerprints. Furthermore, this thesis aims to present the feature importance of these models for potential hypothesis discovery and development regarding the pathological mechanisms of DIFLD. Finally, seven compounds of interest, including some with contradictory results between in vitro and in vivo experiments, were selected to observe how selected models would classify them.

2 Methods

2.1 KNIME

The Konstanz Information Miner (KNIME) is an intuitive, freely available platform which can be used for data preparation, analysis and machine learning (Berthold et al., 2009). For this thesis, KNIME (version 5.4) was utilized to retrieve and curate the data, develop machine learning models, and evaluate their predictive performances.

2.2 Binary datasets

Three binary classification datasets consisting of compounds labeled as DIFLD-positive or DIFLD-negative were employed to construct the models. One dataset was kindly provided by Cronin & Chrysochoou from Liverpool John Moores University (LJMU) in cooperation with the RISK-HUNT3R project. However, since for most of the compounds in this dataset the designated workflow did not find enough available bioactivity data, two other datasets were additionally generated focusing on approved drugs with increased data availability. Lastly, DILIst which is the largest drug list with a binary DILI classification from the Liver Toxicity Knowledge Base (LTKB) (Thakkar et al., 2020), was utilized to also develop a DILI model for comparative analysis. In summary, three binary classification datasets regarding DIFLD and one binary classification dataset regarding DILI were employed. To improve readability, LJMU, FAERS+DILIst, FAERS+FDA and DILIst are used to indicate the four different datasets.

2.2.1 LJMU

As already mentioned, the LJMU dataset was made available by Cronin & Chrysochoou from Liverpool John Moores University. This provided dataset originally contained 1379 DIFLD-positive and -negative compounds with a DTXSID (a unique substance identifier used in the Environmental Protection Agency CompTox Dashboard) and canonical SMILES. The classification of these compounds was based on in-vivo rat and mice studies (Escher et al., 2022; Vrijenhoek et al., 2022), on in-vitro primary human hepatocytes and in-vivo rodent studies (Cotterill et al., 2020), on in-vivo rodent data collected from RepDose, ToxRefDB & HESS DB (Jain et al., 2021) and on two other original sources

(Kubickova & Jacobs, 2023; Lichtenstein et al., 2020). Since the designated workflow uses ChEMBL IDs to retrieve biological fingerprints, the DTXSID were mapped to their corresponding ChEMBL ID using KNIME. After curation and mapping the dataset consisted of 1183 compounds ready for retrieval of biological and substructural fingerprints. As noted earlier, the designated workflow did not find sufficient biological fingerprints for modeling, thus only substructural fingerprints were calculated to encode these compounds. Since substructural fingerprints are based on chemical structure and are not limited to data availability, all 1183 compounds were ultimately used to train the models. This final dataset showed an imbalance towards the number of DIFLD-negative compounds, containing 901 negatives (76.16%) and 282 positives (23.84%).

2.2.2 FAERS+DILIst

To create a DIFLD dataset which contains more approved drugs with increased data availability for biological fingerprints, post-marketing case reports from the FDA Adverse Event Reporting System (FAERS) were curated to assign DIFLD-positive labels to drugs. The FAERS database has already been used by Zhu et al. to annotate drugs on different hepatotoxicity endpoints (X. Zhu & Kruhlak, 2014). Furthermore, Shin et al. employed the frequencies of post-marketing reports on specific hepatic endpoints, including steatosis, from the PharmaPendium database (Shin et al., 2020).

Three reaction terms (Hepatic Steatosis, Steatohepatitis, and Metabolic-Dysfunction associated Steatohepatitis) were selected in the FAERS Public Dashboard (*FAERS Public Dashboard*, 05.05.2025) to retrieve relevant adverse reaction data regarding DIFLD. First, the Case Count by Generic Name table was downloaded, which lists the total number of reported cases regarding each generic drug name. A cut-off of at least 50 cases was chosen to determine DIFLD-positive compounds, excluding biological therapeutics such as monoclonal antibodies. Since more frequently used drugs produce more reports, the line listing table was additionally considered. This table provides information on each case-report regarding the selected reaction terms. These cases were filtered for reports submitted by healthcare professionals. Furthermore, since many reports included multiple, sometimes likely unrelated reactions, cases were screened for those which only contained one of the selected terms. Following, these were further filtered for specific reports that only considered one active ingredient as the assumed cause for the adverse effect.

Ultimately, both generated lists of generic drug names were concatenated, representing the DIFLD-positive class.
Since DIFLD is a subcategory of DILI, DILI-negative drugs should also be DIFLD-negative. Thus, for DIFLD-negative compounds, all suitable DILI-negatives from the publicly available DILIst dataset (Thakkar et al., 2020) were applied as DIFLD-negatives, excluding those which were positive according to the curated FAERS data.
This dataset was then employed to obtain biological fingerprints, leading to a total of 453 encoded compounds which were used for model training. Of these 453 compounds, 248 (55%) were part of the DIFLD-negative class and 205 (45%) of the DIFLD-positive class.

2.2.3 FAERS+FDA

To also construct a binary DIFLD based dataset which is completely independent from already existing DILI datasets, all FDA approved drugs which did not show up as DIFLD-positive in a curation of FAERS were applied as DIFLD-negatives. More specifically, this dataset contained FDA approved compounds which had 10 or less cases of the three selected reaction terms (Hepatic Steatosis, Steatohepatitis, and Metabolic-Dysfunction associated Steatohepatitis) in the FAERS Case Count by Generic Name, labeled as DIFLD-negatives. DIFLD-positives were curated from the FAERS Public Dashboard, as described before.
After retrieval of biological information, utilizing the designated workflow, 200 (21.23%) DIFLD-positives and 742 (78.77%) DIFLD-negatives were encoded for model training.

2.2.4 DILIst

DILIst is the largest binary classification of drugs regarding DILI. DILIst augments the former DILIrank dataset by adding drugs from four literature datasets applying a statistical approach. While DILIrank was solely based on FDA drug-labeling information, the four added DILI datasets employ different strategies, like utilizing clinical evidence, literature, case-registries and specifically FAERS (Thakkar et al., 2020). Since Füzi et al. already used a curated version of the DILIrank dataset for developing a DILI model, employing target and pathway fingerprints, using DILIst will further provide evidence whether this

approach is effective (Füzi et al., 2023). Moreover, it will allow for a comparison to the other models and datasets which are specifically referring to DIFLD.
Because DILIst only provides drug names, the ChEMBL IDs for the drugs were added manually. After retrieval of biological fingerprints, using the designated workflow, a total of 828 drugs were encoded for model training. This final DILIst based dataset was imbalanced towards the number of DILI-positive compounds, consisting of 535 (64.61%) positives and 293 (35.39%) negatives.

2.3 Target fingerprint

For the retrieval of biological fingerprints, the KNIME workflow from Helmke & Ecker, which was originally based on the Path4Drug workflow from Füzi et al., was utilized (**Figure 2**). A table of compound ChEMBL IDs served as input for data extraction. As a first step the designated workflow generates a list of drug-protein interactions using activity data from five different databases: ChEMBL (Zdrazil et al., 2024), IUPHAR/BPS (Harding et al., 2024), DrugBank (Knox et al., 2024), Therapeutic Target Database (Zhou et al., 2024), and PharmGKB (Whirl-Carrillo et al., 2021). Retrieved target information is filtered for human single protein targets and, for targets from ChEMBL, a pChEMBL value of ≥5, which is equivalent to a potency of ≤10 µM, is required for inclusion. pChEMBL is a metric, utilized to indicate the binding strength of a given molecule to a target and is defined as −log10 molar of potency, IC50, XC50, AC50, EC50, Kd or Ki (Allaway et al., 2018). Thus, higher pChEMBL values express stronger molecular activity. However, since binary fingerprints were used for encoding the compounds, all pChEMBL values ≥5 were seen as an interaction, resulting in a value of one in the final binary fingerprint matrix. Every eligible target protein was also mapped to the corresponding Uniprot ID (Helmke & Ecker, 2025), ultimately leading to a list of compound-target pairs encoded as the ChEMBL ID of each drug and the Uniprot ID of each protein.
Since bioactivity data availability was very low for some compounds, the workflow was supplemented with automated target retrieval from the PubChem database (Kim et al., 2022), applying the same principle. This newly generated PubChem target data retrieval workflow (**Figure 1**) was also constructed in KNIME 5.4 and ultimately implemented in the workflow from Helmke & Ecker. The compound input ChEMBL IDs were mapped to their corresponding PubChem CIDs (compound indicators) using the latest Whole Source

Mapping Files from UniChem (*Index of /Pub/Databases/Chembl/UniChem/Data/whole-SourceMapping*). An API call was performed to obtain biological test results from the PubChem assaysummary. The results were filtered for “Activity Outcome” equals “Active” and only tests clearly specifying a Target GeneID were included. Furthermore, the “Activity Name” was restricted to “Potency”, “Ki”, “IC50”, “AC50”, “EC50”, “ID50”, “Kd”, “DC50”, and “ED50”, with a defined “Activity Value [uM]”. Since activity values were given in micromolar (µM), pChEMBL was calculated using the following formula: 6 – log10([µM]). The results were further filtered for a pChEMBL of ≥5 (**Figure 1**) and the corresponding Target GeneIDs were mapped to Uniprot IDs employing the UniProt id-mapping tool (*Retrieve/ID Mapping | UniProt*). Lastly, the retrieved Uniprot IDs were filtered for “Organism” equals “Homo sapiens (Human)” and reviewed UniProtKB entries. Ultimately, as with the other databases, a list with compound ChEMBL IDs and corresponding protein Uniprot IDs was generated and used to supplement the target fingerprint matrix.

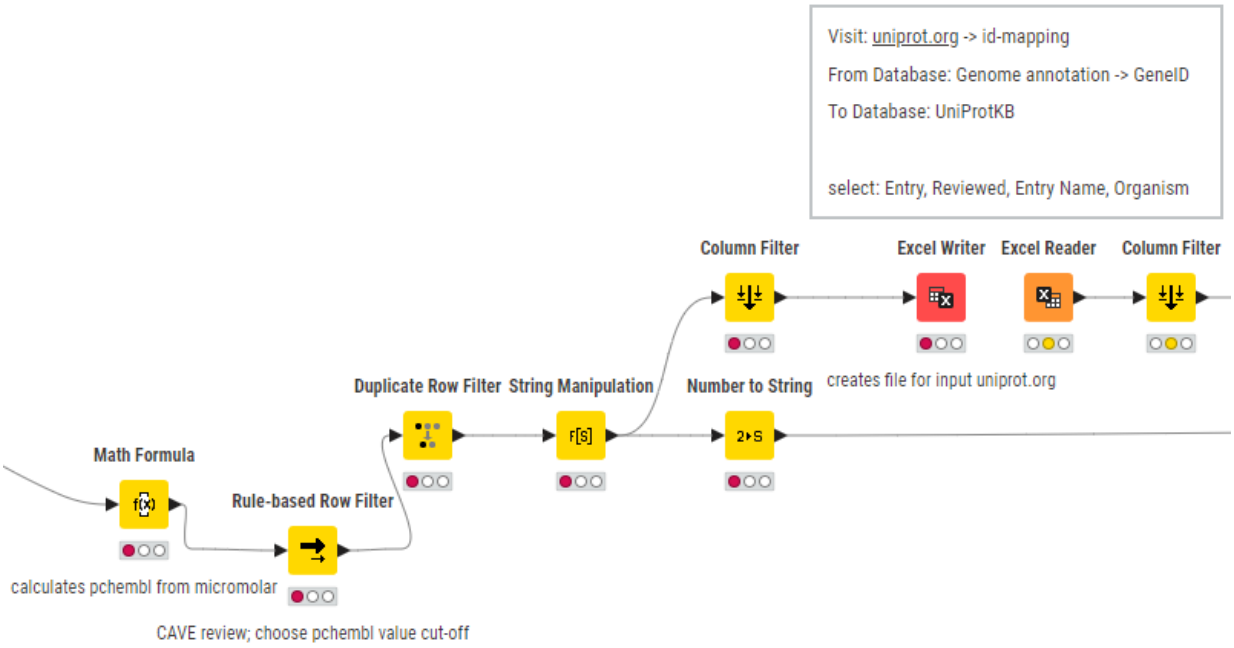


Figure 1. Excerpt of the generated PubChem target data retrieval workflow. The pChEMBL is calculated from retrieved activity values [µM]. Results are filtered for pChEMBL ≥5. Retrieved Target GeneIDs are externally mapped to Uniprot IDs. Finally, Uniprot IDs are filtered for human and reviewed entries.

2.4 Tissue filtering

To refine the list of compound-target pairs for more relevant proteins regarding hepatotoxicity, the tissue-filter from the Path4Drug workflow was employed. This filter uses

expression data from The Tissue Atlas of The Human Protein Atlas (Uhlén et al., 2015) to filter for proteins with sufficient liver expression (Füzi et al., 2021). The utilized expression data contained nTPM (normalized transcripts per million) (Jin et al., 2023) values for each gene-tissue pair. This dataset was restricted to genes with an expression of nTPM ≥11 in liver tissue. Moreover, all target-protein Uniprot IDs were mapped to their corresponding gene name and non-matching proteins were discarded, ultimately resulting in a filtered list of targets with a relevant liver expression level.

2.5 Pathway fingerprint

To obtain relevant compound-pathway connections, the databases MINT (Chatr-aryamontri et al., 2007) and IntAct (del Toro et al., 2022), were used to extract directly interacting proteins of the previously retrieved targets (Helmke & Ecker, 2025). These interactors were filtered for only human, reviewed proteins and utilized in a statistical overrepresentation test, conducted via the Reactome pathway analysis API service (Füzi et al., 2021). The pathway analysis identifies significantly overrepresented pathways through the number of pathway-related proteins in the submitted dataset. Consequently, for each compound-target pair a list of proteins is used to obtain relevant pathways. To decide whether the associated pathway is overrepresented or not, a False-discovery Rate (FDR) of 0.05 served as the cut-off, as done by Füzi et al.

Additionally, all retrieved pathways were filtered for liver expressed pathways by using the Reactome Tissue Distribution analysis tool (Helmke & Ecker, 2025).

After retrieval of liver-relevant targets and pathways of every compound, binary fingerprint matrices were created. While each column header indicated a specific target (Uniprot ID) or pathway (R-HSA ID), row headers represented the compounds (ChEMBL ID) (**Figure 3**). A value of one indicated a retrieved interaction with the target or a connection to the pathway. However, a value of zero indicated that no interaction with the target or no connection to the pathway could be found.

Ultimately, compounds from the original datasets, for which the designated workflow could not identify any targets, were excluded. Since this was a major issue with the LJMU dataset, which mostly consisted of chemicals with low data availability, a biological fingerprint based on targets and pathways could not be applied to this dataset. However, for the other datasets (FAERS+DILIst, FAERS+FDA and DILIst), which were based on

approved drugs, a sufficient number of compounds with at least one target was obtained, with an overall exclusion rate of 30.40%. Moreover, it should be noted that drug exclusion resulted not only from data availability but also from lack of fingerprint uniqueness. Compounds which exhibited the exact same fingerprint as others were excluded, meaning that only unique drugs were ultimately used for modeling. To enable model comparison, this principle was applied to all three fingerprint categories (targets, pathways, bitvectors) independently, resulting in a consistent set of compounds, regardless of descriptor type.

2.6 PubChem bitvectors for structural fingerprint

For structural fingerprints, the PubChem Substructure Fingerprint was utilized, which encodes compounds by an arranged list of 881 binary bits. Each bit determines the presence (1) or absence (0) of certain structural features, such as the count of an element (e.g. >= *2 N*), chemical ring systems (e.g. >= *1 hetero-aromatic ring*), atom-pairs (e.g. *N-N*), atom nearest neighbors, or SMARTS patterns (e.g. *O-C-C=N*) (Helmke & Ecker, 2025; Kim, 2021). PubChem uses this fingerprint and the Tanimoto coefficient for 2-D molecular similarity assessment between two compounds (Kim, 2021). Before calculation of the bitvectors, SMILES of each compound were standardized, utilizing the KNIME-standardizer node (Sosnin, 2024). In this harmonizing process, compounds underwent sanitization, reionization, normalization, charge neutralization, tautomer canonicalization and removal of stereochemistry, using the same in-house standardization pipeline as Helmke & Ecker. Following this, the SMILES were used as input to determine the 881 bitvectors for each compound, employing the KNIME CDK Fingerprints node (Beisken et al., 2013). Ultimately, the list of bits for each compound was split up into a binary matrix, with column headers defining the bit-type (bitvector0-880) and row headers defining the corresponding compounds (ChEMBL ID) (**Figure 3**).

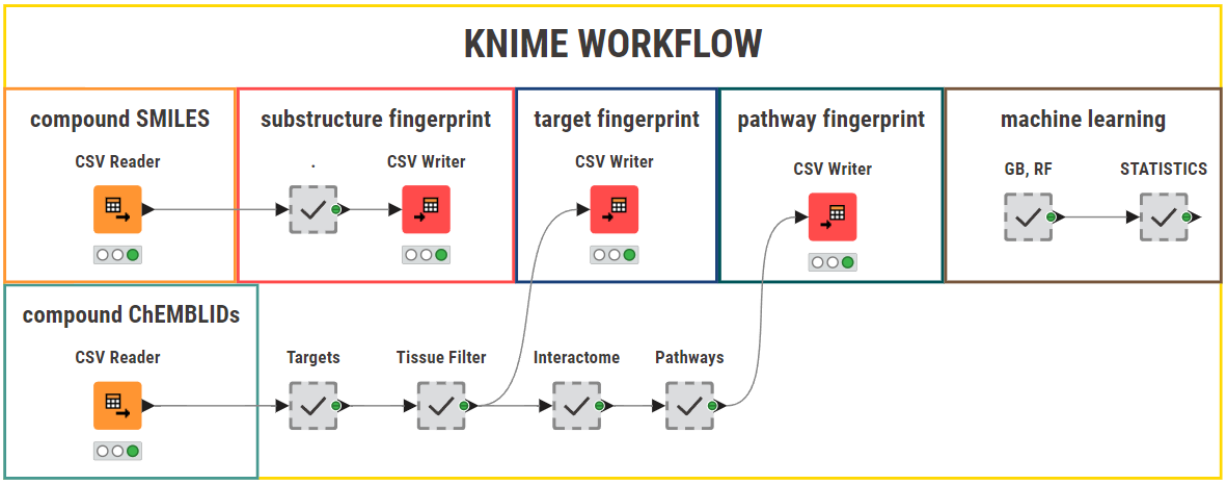


Figure 2. Schematic overview of the designated KNIME workflow. Dataset compounds ChEMBL IDs serve as input for the retrieval of target and pathway interactions. Retrieved targets and pathways with relevant liver expression for each compound are used for generating binary matrices, representing the target and pathway fingerprint. Compounds SMILES are standardized and used for computing the structure-based fingerprint (PubChem BV). Combinations of different fingerprint-matrices of a consistent set of compounds are used for model training. Performance is evaluated through ten-fold cross-validation.

#	RowID	molecule_ChE... *String*	Steatosis *Number (integer)*	P08183 *Number (integer)*	R-HSA-2161522 *Number (integer)*	bitvector353 *Number (integer)*
1	Row0	CHEMBL100116	0	1	1	0
2	Row1	CHEMBL1004	0	0	0	0
3	Row2	CHEMBL1009	0	0	0	0
4	Row3	CHEMBL1014	1	1	1	0
5	Row4	CHEMBL1016	1	1	1	0
6	Row5	CHEMBL1027	0	0	0	1
7	Row6	CHEMBL1029	0	0	0	0
8	Row7	CHEMBL1042	1	0	0	0
9	Row8	CHEMBL1046	0	0	0	0
10	Row9	CHEMBL1059	1	0	0	0

Figure 3. Example table view in KNIME illustrating the combined fingerprint matrix. Each row represents a compound. Compounds are identified by their ChEMBL ID. The steatosis column marks whether compounds are DIFLD-positive (1) or -negative (0). Targets are named by their Uniprot ID, pathways by their R-HSA ID, and PubChem BV by their ID number. 1 indicates the presence of a substructure (PubChem BV) or a database entry (targets, pathways), 0 indicates its absence.

2.7 Machine Learning

For machine learning (ML), the designated workflow utilizes random forest (RF) and gradient-boosted tree (GB) models from the H2O Driverless AI extension in KNIME (*KNIME | H2O.Ai*).

Both, RF and GB, are decision tree based ensemble learning models, which are regarded as robust and high-performing ML techniques for QSAR modeling (Boldini et al., 2023). While RF learns parallelly, by building decision trees from random feature selection ("bagging"), GB works serially, selecting strongly correlated features and integrating them into multiple tree structures ("boosting") (Cha et al., 2021; C.-H. Chen et al., 2020).
Default settings were chosen regarding tree depth (10) and number of models (100), when configurating the model learner KNIME nodes. Moreover, a static seed was employed for the RF learner to enable reproducibility, and a default learning rate of 0.1 was selected for the GB machine learner.
While these ML models determine a probability score for each compound to be DIFLD-positive and negative, they also calculate the absolute and scaled importance of each fingerprint, which generally reflect how valuable a feature was for building the decision trees within the model (C.-H. Chen et al., 2020). A probability score of ≥0.5 for either DIFLD-positivity or DIFLD-negativity was defined as a positive or negative prediction, respectively. The scaling of feature-importance between 0 and 1 is based on the fingerprint ranked as most important (=1).

2.8 Dealing with class imbalances

Since significant class imbalance was present in three out of four datasets, two different approaches were applied to balance the number of positive and negatives compounds. One is known as Synthetic Minority Oversampling (SMOTE), where the minority class is supplemented with artificial samples until both categories are balanced. To be more specific, synthetic examples of minority class compounds are created based on nearest neighbors of real minority class compounds from the original, imbalanced dataset (Hao et al., 2014).
The other balancing method is known as Undersampling (US), where the majority class is reduced into subsets, which are then paired with the minority class compounds, ultimately resulting in one prediction per subset for each drug. These subset predictions were then used to apply a consensus for the final predicted classification of a compound. The number of subsets was determined by the extent of class imbalance. For the LJMU dataset (76.16% negatives) and the FAERS+FDA dataset (78.77% negatives), the majority class was divided into four subsets, as done by Helmke & Ecker, resulting in subsets

which are slightly imbalanced towards the positive compounds (Helmke & Ecker, 2025). For the FAERS+DILIst dataset (55% negatives) and the DILIst dataset (64.61% positives), the majority class was divided into two subsets, resulting in a subset decently imbalanced towards the positive compounds for FAERS+DILIst and in a subset slightly imbalanced towards the negative compounds for DILIst. Although US partially led to imbalanced datasets, the method was applied to all datasets to assess its impact on model performance. The consensus of the predictions for each compound was determined by majority vote of the subset models, with ties resolved using the mean probability score.
Lastly, the third approach was to implement the original imbalanced (Imb) datasets without applying any balancing methods, thus further enabling assessment of their impact on model performance.

2.9 Performance evaluation

A ten-fold cross-validation was employed, which is a common practice for performance evaluation of QSAR-models (Mitchell, 2014). The compounds were split up into 10 different folds. Each fold was then once applied as a test set, while the other folds were combined and used as a training set. Consequently, 10 separate models are generated, resulting in a prediction for each compound from the original dataset (Mitchell, 2014).
As a next step, several performance metrics are calculated, including balanced accuracy (BA), the Matthews correlation coefficient (MCC), sensitivity, specificity, accuracy and precision. These performance indicators are computed using the outcomes of correct and incorrect predictions categorized as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as can be seen from their respective formulas (Duy & Srisongkram, 2025; Helmke & Ecker, 2025):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{MCC} = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$BA = \frac{Sensitivity + Specificity}{2}$$

As outlined by Chicco et al., the MCC is thought to be the superior metric for evaluation of binary classification in ML, outperforming the receiver operating characteristic curve (ROC AUC) (Chicco & Jurman, 2023). Secondly, the BA was chosen because it is known to be more fitting for imbalanced datasets, as it is defined as the arithmetic mean of sensitivity and specificity (Thölke et al., 2023).

2.10 Feature space visualization via traditional t-SNE

For feature space visualization, a new KNIME workflow was constructed, employing the t-SNE node from the KNIME Statistics Nodes (Labs) extension (*T-SNE (L. Jonsson)*), which is based on *T-SNE-Java* from Johnsson (Jonsson, 2014/2025).
T-Distributed stochastic Neighbor Embedding (t-SNE) is a commonly applied dimensionality reduction technique for data space visualization. Through transformation to a two-dimensional map, t-SNE allows for human interpretation of complex, high—dimensional feature spaces, by visualizing clusters and patterns in the data (Orlov et al., 2024). Ultimately, the KNIME workflow (**Figure 4**) was built to generate an interactive scatter plot of the calculated t-SNE dimensions of each compound and to implement a visualization of correct and incorrect classifications (TP, TN, FP, FN). Furthermore, the seven external compounds of interest were also highlighted to see where they are positioned, providing insight into the model's decision-making process and enabling visual validation.

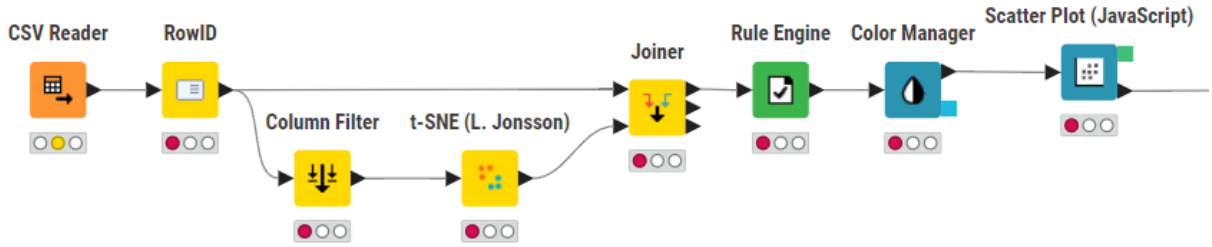


Figure 4. Schematic excerpt of the generated feature visualization workflow. Compounds class, prediction and fingerprint served as input. The t-SNE node determines two dimensions based on each compounds fingerprint. Correct and incorrect predictions are labeled as TP, TN, FP, or FN and color-coded. Finally, all compounds are visualized as data points in a two-dimensional scatter plot.

2.11 Chemical space visualization via MolCompass

A second interactive scatter plot was generated for chemical space visualization, by employing the publicly available MolCompass KNIME node (Sosnin, 2024). MolCompass is primarily based on parametric t-SNE, which utilizes an artificial neural network to depict chemical structures in a two-dimensional space. The main difference between traditional t-SNE and parametric t-SNE is the fact that parametric t-SNE is deterministic. More specifically, while parametric t-SNE projects compounds on a predefined, consistent two-dimensional chemical map, original t-SNE is non-deterministic, leading to different data point arrangements upon every execution. The MolCompass KNIME node utilizes SMILES as an input to compute structure-based ECFPs (Extended-Connectivity Fingerprints), which the chemical map is based on (Sosnin, 2024). Consequently, the scatter plot did not depict the compounds in the same way as perceived by the models, as these were trained on PubChem BV, target, or pathway fingerprints. However, since parametric t-SNE is deterministic, meaning that the same exact compounds always end up at the same exact coordinates, MolCompass enabled direct comparison of the ECFP-based chemical space between the different datasets. As with the feature space, the compounds of interest and the prediction outcome class (TP, TN, FP, FN) were additionally visualized by color.

2.12 Uncertainty and applicability domain

Two probability scores between zero and one were computed for each compound during the ten-fold cross-validation, including P (steatosis =0) and P (steatosis =1), which reflect the predicted likelihood of the compound belonging to the negative and positive class,

respectively. These probability scores were utilized to determine the uncertainty (UC), by calculating the absolute difference between 1 and the higher prediction score for each compound, leading to values between 0 (no UC) and 0.5 (highest UC; same probability score for both classes). This was done for every compound to evaluate how excluding those with an UC of ≥0.35 would influence model performance, following Helmke & Ecker. Furthermore, a fixed cutoff of 0.35 was also applied to the seven external compounds of interest to classify their predictions as “UC_Risk” or “UC_OK”.

To evaluate whether the external compounds of interest were part of the model’s applicability domain (AD), the distance of each test compound to its nearest neighbor in the training set was considered. For this task, another new KNIME workflow (**Figure 5**) was generated, which at first calculates the Tanimoto distance between every pair of compounds in the dataset. The Tanimoto coefficient is commonly utilized to calculate similarities between fingerprints of molecules (Bajusz et al., 2015). However, the Tanimoto distance (or Soergel distance) between two molecules is defined as the complement of their Tanimoto coefficient, with values ranging from 0 (identical) to 1 (no shared features) (Bajusz et al., 2015). Since the Tanimoto coefficient only considers the bits set to 1 in both compounds’ fingerprints, it is especially suitable for comparing molecular fingerprints, which contain relatively few active (1) bits, such as ECFP4 (Tripp et al., 2023). Applying this rationale, the Tanimoto distance was employed for distance-to-nearest Neighbor (DtNN) calculation, particularly because PubChem bitvectors were used as fingerprints and the PubChem database itself applies the Tanimoto coefficient to this fingerprint for molecular similarity assessment, as mentioned before (Kim, 2021). Although others used a Euclidean distance-based k-nearest neighbors algorithm (Duy & Srisongkram, 2025; Tropsha et al., 2003), applying the Tanimoto distance appeared appropriate, since the target fingerprint also contained a low number of active (1) bits. After calculating the Tanimoto distances between all compound pairs, the DtNN for each compound is determined. Moreover, only regarding distances between training compounds (**Figure 6**), the mean and the SD (standard deviation) of DtNN is computed to determine a cutoff value, as done by Duy & Srisongkram and Tropsha et al. The cutoff was applied to the Tanimoto distance of each external test compound to their nearest training compound neighbor. The cutoff works as follows (Duy & Srisongkram, 2025):

Test compound within applicability domain:

$$DtNN_{test} < Mean(DtNN_{train}) + Z \times SD(DtNN_{train})$$

Test compound outside applicability domain:

$$DtNN_{test} \geq Mean(DtNN_{train}) + Z \times SD(DtNN_{train})$$

The Z-score determines the significance level and is defined empirically. Since it is usually set to 0.5 (Duy & Srisongkram, 2025; Tropsha et al., 2003), this value was also chosen for this work. Ultimately, the threshold classified each external test compound as “in_domain” or “outside_domain”.

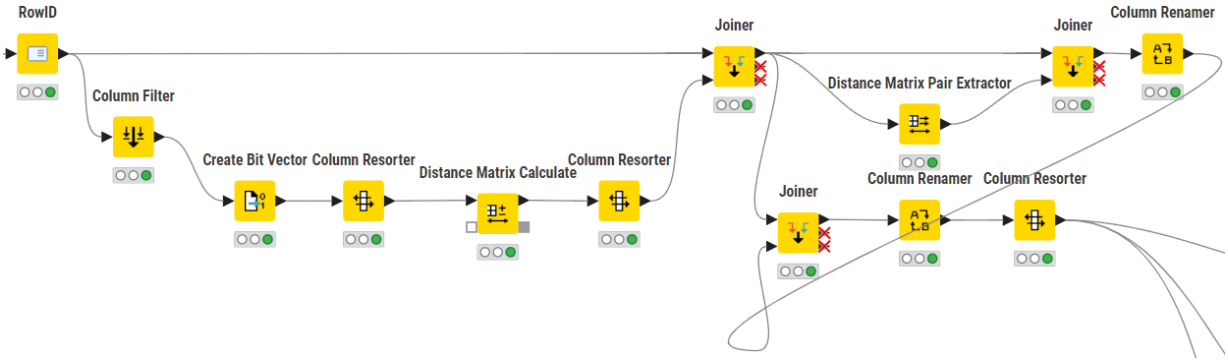


Figure 5. Excerpt of the generated applicability domain workflow. DtNN analysis was applied to define the models applicability domain. Compounds are labeled as Test and Training molecules. Each fingerprint is used to create a singular Bit Vector. The Tanimoto distance is calculated and extracted for each molecule pair. The shortest distances between train-train and test-train pairs for each compound is determined. The DtNN of each training molecule is utilized to calculate the cut-off. The DtNN of each test compound is employed for classification as “in_domain” or “outside_domain”.

#	RowID	Object1 *String*	SetType_Object1 *String*	Object2 *String*	SetType_Object2 *String*	Distance *Number (double)*
1	Row38	CHEMBL2	Train	CHEMBL707	Train	0.254
2	Row35	CHEMBL2	Train	CHEMBL611	Train	0.257
3	Row22	CHEMBL2	Train	CHEMBL2105760	Train	0.43
4	Row39	CHEMBL2	Train	CHEMBL76370	Train	0.431
5	Row24	CHEMBL2	Train	CHEMBL3349607	Train	0.442
6	Row35	CHEMBL2	Train	CHEMBL623	Train	0.444
7	Row29	CHEMBL2	Train	CHEMBL45816	Train	0.458
8	Row22	CHEMBL2	Train	CHEMBL23	Train	0.484
9	Row31	CHEMBL2	Train	CHEMBL49	Train	0.495
10	Row23	CHEMBL2	Train	CHEMBL24778	Train	0.496

Figure 6. Table view in KNIME illustrating the Tanimoto distances between training compounds. Compounds are identified by their ChEMBL ID. Distances for train-train pairs are sorted in ascending order. The shortest distance for each compound is defined as DtNN and used to compute the applicability domain cut-off.

3 Results

A total of four binary classification datasets (**Table 1**) were used in this work, including three datasets regarding DIFLD and one comparative dataset regarding DILI in general. Only unique drugs regarding all fingerprint types (PubChem BV, targets, pathways) were used for the FAERS+DILIst, FAERS+FDA and DILIst datasets, and only unique compounds concerning PubChem BV were employed for the LJMU dataset.

The LJMU dataset was mainly based on in-vitro or in-vivo data and comprised 901 (76.16%) DIFLD-negatives and 282 (23.84%) positives. The FAERS+DILIst dataset was generated using case reports from FAERS (*FAERS Public Dashboard*, 05.05.2025) for positives and the DILIst dataset (Thakkar et al., 2020) for negatives, leading to 248 (55%) DIFLD-negatives and 205 (45%) positives. The FAERS+FDA dataset was also constructed utilizing case reports from FAERS and a list of all FDA approved drugs, resulting in 742 (78.77%) DIFLD-negatives and 200 (21.23%) positives. Lastly, DILIst, the largest binary classification of drugs regarding DILI (Thakkar et al., 2020), was employed as a fourth, comparative dataset, which ultimately included 293 (35.39%) DILI-negatives and 535 (64.61%) positives.

Table 1. Sizes and class balances of datasets.

Dataset	Unique comp.	Positives	Negatives
LJMU	1183	282 (23.84%)	901 (76.16%)
FAERS+DILIst	453	205 (45.00%)	248 (55.00%)
FAERS+FDA	942	200 (21.23%)	742 (78.77%)
DILIst	828	535 (64.61%)	293 (35.39%)

Unique comp. = compounds with a unique fingerprint.

3.1 Fingerprint length

While the number of the biological fingerprints (targets, pathways) differed throughout the datasets, the length of the structural fingerprint (PubChem BV) was consistent (**Table 2**). Since only PubChem BV were used for encoding the LJMU dataset compounds, the fingerprint length for this dataset was 881. For the FAERS+DILIst dataset, 849 targets with sufficient liver expression and 1977 liver-expressed pathways were part of the biological fingerprint. The FAERS+FDA dataset led to the longest fingerprint with 1105 targets and 2024 pathways. Finally, the DILIst datasets fingerprint included 997 targets and 2003

pathways. Unsurprisingly, larger datasets with more drugs generated a higher fingerprint count.

Table 2. Fingerprint lengths of datasets

Dataset	Unique comp.	Unique PubChem BV	Unique targets	Unique pathways
LJMU	1183	881	-	-
FAERS+DILIst	453	881	849	1977
FAERS+FDA	942	881	1105	2024
DILIst	828	-	997	2003

Unique comp. = compounds with a unique fingerprint; *PubChem BV* = PubChem Bitvectors (substructural fingerprint).

3.2 Comparison of the FAERS-derived DIFLD-positives with DILI-positives

In an attempt to assess the validity of the positive class in the FAERS+DILIst dataset, DILI-positive drugs from DILIst (Thakkar et al., 2020) were compared with compounds identified as steatogenic in the curation of FAERS (**Figure 7**). The FAERS+DILIst dataset itself was generated by combining FAERS-derived positives with DILI-negative drugs from DILIst.

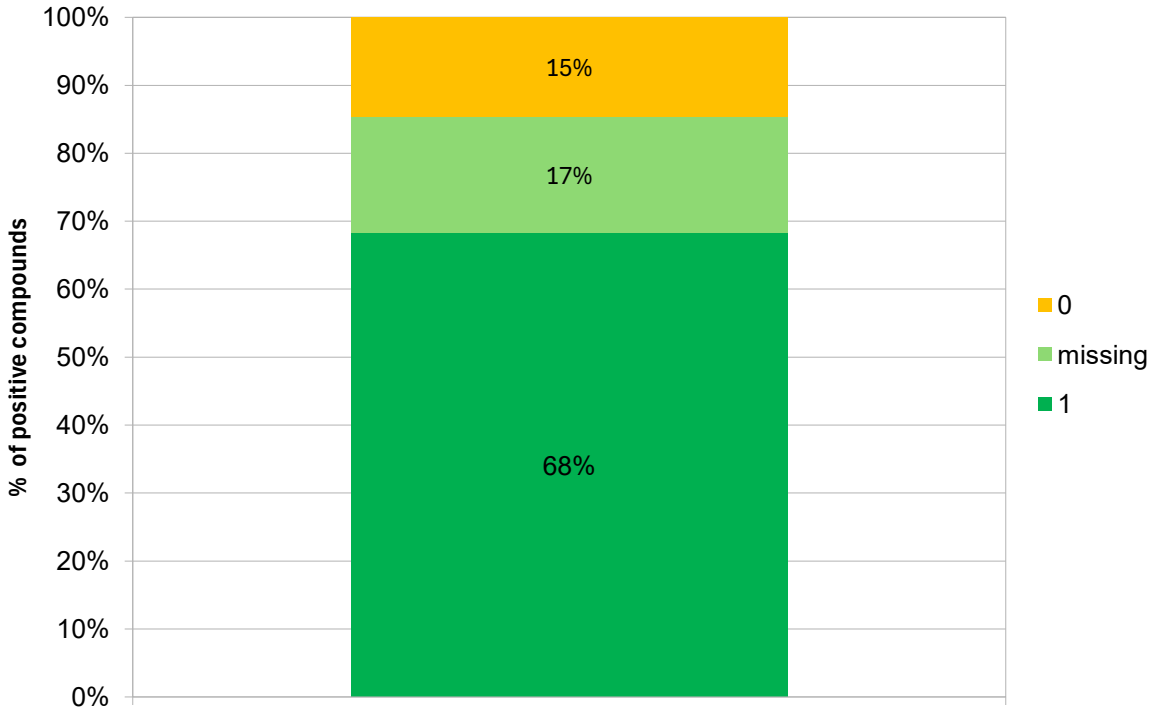


Figure 7. Bar chart illustrating the comparison of FAERS-derived DIFLD-positives with DILI-positives as reported in the DILIst dataset (Thakkar et al., 2020). 140 (68%) of DIFLD-positives are also classified as DILI-positive (1). 35 (17%) of DIFLD-positives are missing in the DILIst dataset. 30 (15%) of DIFLD-positives are contradictory labeled as DILI-negatives (0).

From the 205 DIFLD-positive compounds regarding the curation of FAERS, 140 (68%) were also classified as DILI-positive in the DILIst dataset. However, 35 (17%) of the DIFLD-positive assigned drugs could not be found in the DILIst dataset, and 30 (15%) were part of the DILI-negatives. Interestingly, among the 30 contradicting compounds, several opioids such as fentanyl, oxycodone and codeine were found, whereas none were present within the 140 concordant drugs. Moreover, when examining the 35 DIFLD-positive assigned compounds, which were missing in the DILIst dataset, a relatively high number of tyrosine kinase inhibitors, including Ruxoltinib, Upadacitinib and Alectinib, was observed. However, among the 140 concordant drugs, there were also several drugs of this class, including Imatinib, Sunitinib and Dasatinib.

3.2 Model performances

Four different binary classification datasets (LJMU, FAERS+DILIst, FAERS+FDA, DILIst) were used for modeling (**Table 1**). Compounds were encoded via substructural (PubChem BV) and biological (targets, pathways) fingerprint-types. Since class imbalance was present in all four datasets, two different balancing techniques, US and SMOTE, were applied. However, the entire imbalanced datasets (Imb) were also modeled without any balancing. Two different types of tree-based ensemble ML models, RF and GB were employed for each dataset. A ten-fold cross-validation was used for performance evaluation, ultimately calculating the BA, MCC, sensitivity, specificity, accuracy and precision. Models were ranked from highest to lowest BA.

Table 3. Ten-fold cross-validation results of the LJMU dataset models

Fingerprint	Model	BA	MCC	Sensitivity	Specificity	Accuracy	Precision
BV	SMOTE_GB	0.627	0.242	0.472	0.782	0.708	0.404
BV	US_GB	0.608	0.185	0.706	0.511	0.557	0.311
BV	US_RF	0.605	0.180	0.706	0.505	0.553	0.309
BV	Imb_GB	0.598	0.209	0.351	0.846	0.728	0.416
BV	SMOTE_RF	0.581	0.168	0.337	0.825	0.708	0.375
BV	Imb_RF	0.542	0.151	0.124	0.960	0.761	0.493

BA = balanced accuracy; *MCC* = Matthews correlation coefficient; *BV* = PubChem Bitvectors (substructural fingerprint); *SMOTE* = synthetic minority oversampling technique; *US* = undersampling; *Imb* = imbalanced (no balancing technique applied).

For the LJMU dataset only PubChem BV were employed as the designated workflow did not retrieve sufficient bioactivity data for these compounds.

Regarding results, GB outperformed RF independent of balancing technique, with SMOTE_GB yielding the highest BA (0.63), MCC (0.24) and specificity (0.78). Unsurprisingly, US scored the highest in sensitivity (0.71), however this came at the cost of specificity and precision (**Table 3**).

Table 4. Ten-fold cross-validation results of the FAERS+DILIst dataset models

Fingerprint	Model	BA	MCC	Sensitivity	Specificity	Accuracy	Precision
TARGETS	US_GB	0.707	0.415	0.785	0.629	0.700	0.636
TARGETS	SMOTE_GB	0.700	0.413	0.580	0.819	0.711	0.726
TARGETS_PATHS	US_RF	0.697	0.403	0.829	0.565	0.684	0.612
BV_PATHS	Imb_RF	0.693	0.399	0.576	0.810	0.704	0.715
TARGETS	SMOTE_RF	0.687	0.385	0.580	0.794	0.698	0.700
ALL	SMOTE_RF	0.686	0.374	0.634	0.738	0.691	0.667
TARGETS_BV	Imb_GB	0.684	0.381	0.566	0.802	0.695	0.703
ALL	US_RF	0.683	0.379	0.834	0.532	0.669	0.596
TARGETS	Imb_GB	0.682	0.382	0.546	0.819	0.695	0.713
TARGETS_BV	SMOTE_GB	0.678	0.372	0.546	0.810	0.691	0.704
TARGETS_BV	US_GB	0.676	0.358	0.795	0.556	0.664	0.597
BV_PATHS	US_RF	0.675	0.362	0.829	0.520	0.660	0.588
PATHS	US_RF	0.674	0.355	0.800	0.548	0.662	0.594
TARGETS_PATHS	SMOTE_RF	0.673	0.348	0.624	0.722	0.678	0.650
TARGETS_PATHS	US_GB	0.671	0.347	0.785	0.556	0.660	0.594
TARGETS	US_RF	0.670	0.359	0.849	0.492	0.653	0.580
ALL	Imb_RF	0.669	0.341	0.595	0.742	0.675	0.656
TARGETS_BV	SMOTE_RF	0.669	0.341	0.595	0.742	0.675	0.656
ALL	US_GB	0.667	0.341	0.790	0.544	0.656	0.589
TARGETS	Imb_RF	0.662	0.340	0.517	0.806	0.675	0.688
BV_PATHS	SMOTE_RF	0.662	0.325	0.610	0.714	0.667	0.638
PATHS	US_GB	0.662	0.331	0.795	0.528	0.649	0.582
BV_PATHS	Imb_GB	0.654	0.326	0.502	0.806	0.669	0.682
PATHS	SMOTE_RF	0.654	0.308	0.610	0.698	0.658	0.625
BV_PATHS	SMOTE_GB	0.648	0.307	0.517	0.778	0.660	0.658
BV_PATHS	US_GB	0.643	0.293	0.771	0.516	0.631	0.568
ALL	SMOTE_GB	0.642	0.298	0.502	0.782	0.656	0.656
TARGETS_PATHS	Imb_RF	0.642	0.287	0.571	0.714	0.649	0.622
TARGETS_PATHS	Imb_GB	0.642	0.298	0.498	0.786	0.656	0.658
PATHS	Imb_GB	0.638	0.284	0.522	0.754	0.649	0.637
ALL	Imb_GB	0.637	0.284	0.507	0.766	0.649	0.642
TARGETS_PATHS	SMOTE_GB	0.636	0.284	0.502	0.770	0.649	0.644
TARGETS_BV	US_RF	0.632	0.286	0.844	0.419	0.611	0.546

PATHS	Imb_RF	0.630	0.263	0.541	0.718	0.638	0.613
TARGETS_BV	Imb_RF	0.630	0.266	0.517	0.742	0.640	0.624
PATHS	SMOTE_GB	0.616	0.242	0.483	0.750	0.629	0.615
BV	US_RF	0.615	0.261	0.863	0.367	0.592	0.530
BV	SMOTE_GB	0.611	0.232	0.463	0.758	0.625	0.613
BV	Imb_GB	0.606	0.223	0.459	0.754	0.620	0.606
BV	US_GB	0.602	0.214	0.776	0.427	0.585	0.528
BV	SMOTE_RF	0.589	0.180	0.517	0.661	0.596	0.558
BV	Imb_RF	0.577	0.158	0.473	0.681	0.587	0.551

BA = balanced accuracy; *MCC* = Matthews correlation coefficient; *BV* = PubChem Bitvectors (substructural fingerprint); *SMOTE* = synthetic minority oversampling technique; *US* = undersampling; *Imb* = imbalanced (no balancing technique applied).

Regarding the FAERS+DILIst dataset, the target-based fingerprint achieved the best results, with a BA of 0.71, an MCC of 0.42 and a sensitivity of 0.79 for the US_GB model, which was also the highest scoring model throughout all datasets. Other similar scoring fingerprints were mostly combinations that included targets. Interestingly, although the pathway and PubChem BV fingerprints scored notably worse individually, their combination yielded results similar to those of the target-based fingerprints. Both RF and GB appeared among the best scoring models in this dataset (**Table 4**).

Table 5. Ten-fold cross-validation results of the FAERS+FDA dataset models

Fingerprint	Model	BA	MCC	Sensitivity	Specificity	Accuracy	Precision
TARGETS	US_GB	0.661	0.269	0.655	0.667	0.665	0.347
TARGETS	US_RF	0.650	0.255	0.605	0.695	0.676	0.349
BV_PATHS	US_RF	0.650	0.249	0.650	0.650	0.650	0.333
TARGETS_PATHS	US_RF	0.649	0.248	0.650	0.648	0.649	0.332
TARGETS_BV	US_RF	0.648	0.244	0.655	0.640	0.643	0.329
TARGETS_BV	US_GB	0.647	0.242	0.670	0.624	0.634	0.324
ALL	US_RF	0.646	0.243	0.640	0.652	0.650	0.332
PATHS	US_RF	0.641	0.234	0.635	0.647	0.644	0.326
ALL	US_GB	0.639	0.230	0.645	0.633	0.636	0.322
BV_PATHS	US_GB	0.637	0.227	0.655	0.620	0.627	0.317
TARGETS_PATHS	US_GB	0.635	0.223	0.655	0.616	0.624	0.315
TARGETS	SMOTE_GB	0.633	0.291	0.380	0.887	0.779	0.475
TARGETS_BV	SMOTE_RF	0.624	0.282	0.350	0.899	0.782	0.483
TARGETS	SMOTE_RF	0.618	0.243	0.385	0.852	0.753	0.412
PATHS	US_GB	0.617	0.192	0.655	0.580	0.596	0.296
TARGETS	Imb_GB	0.616	0.296	0.300	0.933	0.798	0.545
TARGETS_PATHS	SMOTE_RF	0.607	0.218	0.370	0.844	0.743	0.389
BV_PATHS	SMOTE_RF	0.603	0.225	0.330	0.876	0.760	0.418

ALL	SMOTE_RF	0.601	0.231	0.310	0.892	0.769	0.437
ALL	SMOTE_GB	0.590	0.237	0.250	0.930	0.786	0.490
TARGETS_BV	SMOTE_GB	0.590	0.215	0.275	0.904	0.771	0.437
TARGETS_BV	Imb_GB	0.589	0.237	0.245	0.933	0.787	0.495
ALL	Imb_GB	0.587	0.247	0.230	0.945	0.793	0.529
PATHS	SMOTE_GB	0.583	0.196	0.270	0.896	0.763	0.412
BV_PATHS	SMOTE_GB	0.582	0.223	0.230	0.934	0.785	0.484
TARGETS_PATHS	SMOTE_GB	0.577	0.193	0.245	0.910	0.769	0.422
BV_PATHS	Imb_GB	0.570	0.219	0.185	0.956	0.792	0.529
TARGETS_PATHS	Imb_GB	0.566	0.184	0.200	0.933	0.777	0.444
PATHS	SMOTE_RF	0.566	0.135	0.300	0.832	0.719	0.324
PATHS	Imb_RF	0.558	0.213	0.145	0.972	0.796	0.580
PATHS	Imb_GB	0.556	0.151	0.190	0.922	0.766	0.396
BV	Imb_GB	0.552	0.140	0.185	0.919	0.763	0.381
BV	SMOTE_GB	0.548	0.114	0.215	0.881	0.740	0.328
BV	US_RF	0.547	0.077	0.595	0.499	0.519	0.242
BV_PATHS	Imb_RF	0.542	0.165	0.110	0.973	0.790	0.524
BV	SMOTE_RF	0.540	0.082	0.265	0.815	0.699	0.279
BV	US_GB	0.536	0.059	0.580	0.492	0.511	0.235
BV	Imb_RF	0.530	0.118	0.090	0.969	0.782	0.439
TARGETS_PATHS	Imb_RF	0.528	0.118	0.085	0.972	0.783	0.447
TARGETS_BV	Imb_RF	0.528	0.140	0.070	0.985	0.791	0.560
TARGETS	Imb_RF	0.523	0.126	0.060	0.987	0.790	0.545
ALL	Imb_RF	0.523	0.113	0.065	0.981	0.787	0.481

BA = balanced accuracy; *MCC* = Matthews correlation coefficient; *BV* = PubChem Bitvectors (substructural fingerprint); *SMOTE* = synthetic minority oversampling technique; *US* = undersampling; *Imb* = imbalanced (no balancing technique applied).

The target fingerprint-based model was also the best scoring one regarding the FAERS+FDA dataset, which was generated through curation of FAERS and a list of all FDA approved drugs. Since the positive class was constructed in a similar fashion, a certain correspondence in results was expected. However, overall performance was distinctly worse with a BA of 0.66 and an MCC of 0.27 for the highest scoring model. Concerning models based solely on the PubChem BV fingerprint, the ten-fold cross-validation again yielded considerably poorer performances than models based on biological fingerprints (**Table 5**).

As the FAERS+FDA dataset was notably imbalanced (79% negatives vs. 21% positives), the US technique consistently outperformed SMOTE and Imb. Moreover, since US led to four almost balanced subsets, sensitivity and specificity showed very similar results, regarding this balancing approach. Also because of the unbalanced nature of this dataset,

models using the Imb_RF configuration exhibited an extremely low sensitivity close to zero.

Table 6. Ten-fold cross-validation results of the DILIst dataset models

Fingerprint	Model	BA	MCC	Sensitivity	Specificity	Accuracy	Precision
PATHS	US_RF	0.589	0.172	0.492	0.686	0.561	0.740
TARGETS	SMOTE_RF	0.588	0.170	0.518	0.659	0.568	0.734
TARGETS+PATHS	US_GB	0.581	0.156	0.535	0.628	0.568	0.723
TARGETS	US_RF	0.578	0.159	0.381	0.775	0.521	0.755
PATHS	SMOTE_RF	0.577	0.150	0.619	0.536	0.590	0.708
TARGETS	US_GB	0.577	0.147	0.533	0.621	0.564	0.719
TARGETS+PATHS	US_RF	0.569	0.133	0.478	0.659	0.542	0.718
TARGETS	SMOTE_GB	0.564	0.123	0.568	0.560	0.565	0.701
PATHS	US_GB	0.558	0.111	0.525	0.590	0.548	0.700
TARGETS+PATHS	SMOTE_GB	0.554	0.105	0.630	0.478	0.576	0.687
TARGETS	Imb_GB	0.552	0.101	0.649	0.454	0.580	0.684
TARGETS+PATHS	SMOTE_RF	0.551	0.098	0.610	0.491	0.568	0.686
TARGETS+PATHS	Imb_GB	0.534	0.067	0.651	0.416	0.568	0.670
PATHS	SMOTE_GB	0.528	0.055	0.610	0.447	0.552	0.667
PATHS	Imb_GB	0.528	0.055	0.666	0.389	0.568	0.665
TARGETS	Imb_RF	0.508	0.038	0.959	0.058	0.639	0.649
PATHS	Imb_RF	0.504	0.015	0.946	0.061	0.632	0.647
TARGETS+PATHS	Imb_RF	0.499	0.005	0.957	0.041	0.632	0.645

BA = balanced accuracy; *MCC* = Matthews correlation coefficient; *SMOTE* = synthetic minority oversampling technique; *US* = under-sampling; *Imb* = imbalanced (no balancing technique applied).

Interestingly, regarding the largest binary DILI classification dataset DILIst (Thakkar et al., 2020), model performance was significantly worse. Since this dataset was only employed for comparative analysis of the biological fingerprint-based models, PubChem BV were not included in these models. Unlike the two DIFLD-based datasets, which were generated through curation of FAERS, the best performing fingerprint were liver-expressed pathways alone. However, the target fingerprint and combinations of the two scored very similar with a BA of around 0.59 and an MCC of 0.17. As the FAERS+FDA dataset (**Table 5**), the DILIst dataset was also considerably imbalanced (35% negatives and 65% positives), thus models with the Imb_RF configuration likewise showed very low sensitivities (**Table 6**), approaching zero.

3.3 Number and Distribution of active Fingerprints per compound

To analyze the number and distribution of active liver-expressed targets and pathways per compound in each dataset, a boxplot (**Figure 8**) was generated. Interestingly, the best performing FAERS+DILI dataset showed the largest discrepancy in active targets between positive and negative compounds, with a median of 12 active targets per drug for the positive class and 6 for the negative class (**Table 7**). This difference between the classes was present in all three models that employed biological fingerprints, however, for the poorest performing dataset DILIst, the discrepancy was notably smaller, with a median of 9 active targets per positive compound and 6 per negative compound. Regarding pathways, this difference was also present, although the overall numbers of active pathways per compound were much higher. Since this observation raised the question of whether the models classified compounds according to this difference in actives-count, results from the ten-fold cross-validation were filtered for positive compounds with fewer than the median number of actives per drug and negative compounds with more than the median number of actives per drug. This was done for the target based FAERS+DILIst US_GB model and showed a BA of 0.5 (**Table 8**). The overall BA for this model was 0.71, indicating that these compounds were frequently misclassified. Moreover, the results were also filtered for drugs with more than the median number of actives per compound, regardless of class. As expected, this led to a high sensitivity (0.88) and a BA of 0.65. Even though the specificity was low regarding these compounds (0.42), some of the negatives could still be correctly classified, indicating that the count of liver-expressed active targets per drug was not the only factor influencing the model's decision-making.

Table 7. Distribution of active target-count per compound across positives and negatives for the FAERS+DILIst dataset

FAERS+DILIst Dataset	Positives and negatives	Positives	Negatives
Maximum	31.5	39.5	23
Q3	15	20	11
Median	9	12	6
Q1	4	7	3
Minimum	1	1	1

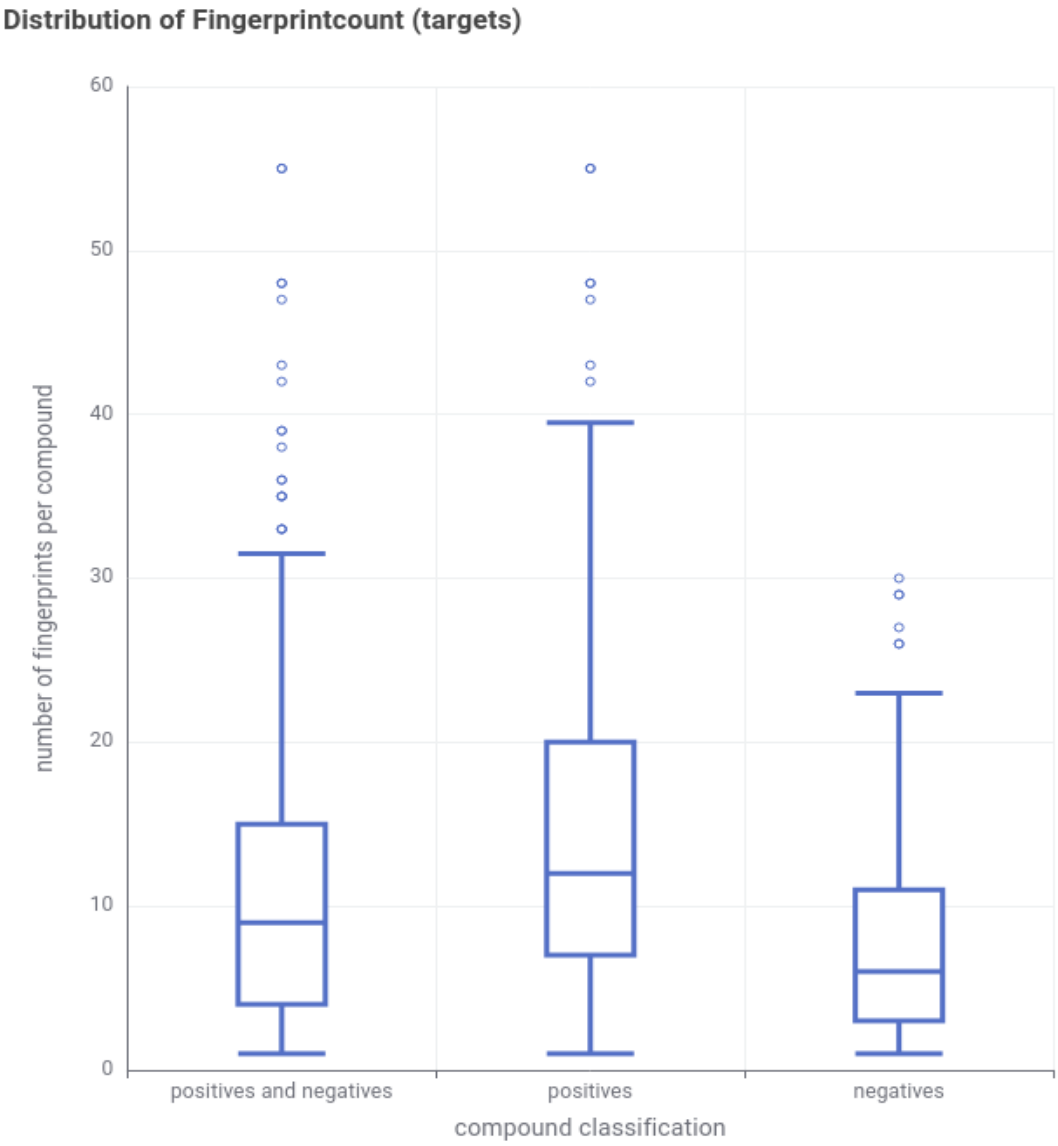


Figure 8. Boxplot illustrating the distribution of the active target-count per compound across the positive and negative class for the FAERS+DILIst dataset. Maximum, Q3, median and Q1 values were higher for DIFLD-positive compounds (**Table 7**).

Table 8. Impact of active-target count-based compound filtering on cross-validation results of the FAERS+DILIst US_GB model

All compounds		
BA	**Specificity**	**Sensitivity**
0.71	0.63	0.79

Negatives with > Median number of active targets per compound Positives with < Median number of active targets per compound		
Confusion Matrix	**Prediction = negative**	**Prediction = positive**
DIFLD negatives	36	49
DIFLD positives	29	39
BA	**Specificity**	**Sensitivity**
0.50	0.42	0.57

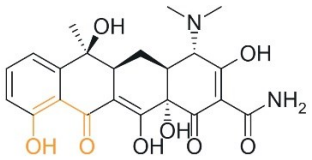
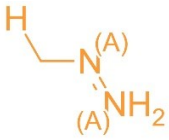
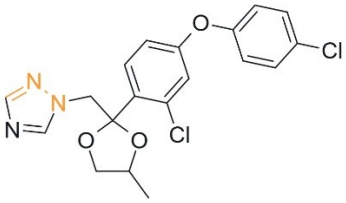
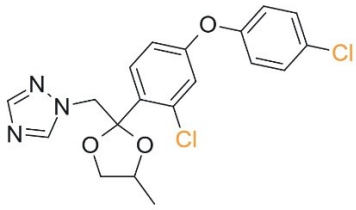
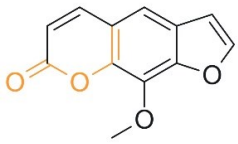
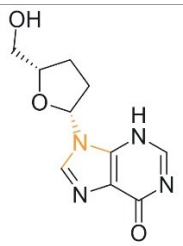
All compounds with > Median number of active targets per compound		
Confusion Matrix	**Prediction = negative**	**Prediction = positive**
DIFLD negatives	36	49
DIFLD positives	17	128
BA	**Specificity**	**Sensitivity**
0.65	0.42	0.88

BA = balanced accuracy.

3.4 Feature Importance and distribution of actives

The scaled importance of each fingerprint (PubChem BV, targets, pathways) was calculated regarding every model. Since the US approach and the ten-fold cross validation generated several subsets, the mean scaled importance for each fingerprint was computed and selected as the final scaled importance. Moreover, the ratio of active compounds per fingerprint was determined for both classes separately. The absolute difference between these ratios was then calculated for each feature to illustrate the discrepancy in occurrence of actives. Ultimately, tables listing all used fingerprints were generated for each model and sorted by the scaled feature importance. The 5 most important substructural features regarding the LJMU dataset-based model which only employed PubChem BV, are displayed in **Table 9**, including one selected example compound from the positive class per fingerprint. Moreover, the top 30 ranked fingerprints regarding models which also used biological features are presented in **Table 10-13**.

Table 9. Top 5 most important features of the LJMU SMOTE_GB model (PubChem BV)

Feature ID	BV substructure	Example compounds (+)	DIFLD+	Ratio in+	DIFLD-	Ratio in-
BV625		 Tetracycline	14	0.050	21	0.023
BV496		 Difenoconazole	29	0.103	16	0.018
BV38	*>= 2 Cl*	 Difenoconazole	62	0.220	93	0.103
BV622		 Methoxsalen	5	0.018	15	0.017
BV391	*N(~C)(~C)(~C)*	 Didanosine	49	0.174	131	0.145

BV = PubChem Bitvectors (substructural fingerprint); *DIFLD+ / DIFLD-* = total count of DIFLD-positive or -negative compounds containing the given substructure; *Ratio in+ / Ratio in-* = fraction of compounds within the respective group containing the given substructure.

Regarding the LJMU dataset-based model, which only used substructural information as fingerprints, three simple SMARTS patterns, one element count, and one simple atom

nearest neighbor pattern accounted for the top five most important features (**Table 9**). Simple SMARTS patterns determine the presence or absence of certain substructures with a specific bond order and bond aromaticity, regardless of count (*List_fingerprints.Pdf*). Tetracycline was selected as a DIFLD-positive example compound, exhibiting the highest ranked structural pattern (*O=C-C:C-O*). For the next two most important features (*N:N-C-[#1]* and *>= 2 Cl*), difenoconazole was chosen to illustrate them, as it contained both. These features were notably overrepresented in the positive class, with ratios of 0.103 and 0.220, respectively. Methoxsalen and didanosine, both compounds that are also positive regarding DILI in the DILIst dataset (Thakkar et al., 2020), represent the remaining two top ranked fingerprints which included another SMARTS substructure (*O=C-O-C:C*) and an atom nearest neighbor pattern (*N(~C)(~C)(~C)*).

Table 10. Top 30 most important features of the FAERS+DILIst SMOTE_RF model (all fingerprint-types)

Feature ID	Feature type	Description	DIFLD+	Ratio in+	DIFLD-	Ratio in-
R-HSA-196849	pathway	Metabolism of water-soluble vitamins and cofactors	63	0.307	14	0.056
R-HSA-9725554	pathway	Differentiation of Keratinocytes	66	0.322	12	0.048
R-HSA-879415	pathway	Advanced glycosylation end-product receptor signaling	141	0.688	82	0.331
R-HSA-2161517	pathway	Abacavir transmembrane transport	131	0.639	70	0.282
R-HSA-9754706	pathway	Atorvastatin ADME	176	0.859	137	0.552
R-HSA-8939246	pathway	RUNX1 regulates transcription of genes	131	0.639	72	0.290
P02768	target	Albumin	94	0.459	47	0.190
R-HSA-189483	pathway	Heme degradation	98	0.478	34	0.137
R-HSA-189445	pathway	Metabolism of porphyrins	98	0.478	34	0.137
R-HSA-9013408	pathway	RHOG GTPase cycle	82	0.400	27	0.109
R-HSA-9753281	pathway	Paracetamol ADME	102	0.498	44	0.177
R-HSA-4570464	pathway	SUMOylation of RNA binding proteins	129	0.629	83	0.335
R-HSA-8878166	pathway	Transcriptional regulation by RUNX2	149	0.727	104	0.419
Q9UNQ0	target	ATP-binding cassette transporter ABCG2	53	0.259	9	0.036
R-HSA-9757110	pathway	Prednisone ADME	170	0.829	130	0.524
R-HSA-1660661	pathway	Sphingolipid de novo biosynthesis	91	0.444	36	0.145

P08183	target	ATP-dependent translocase ABCB1	116	0.566	55	0.222
P10632	target	CYP2C8	70	0.341	24	0.097
R-HSA-917937	pathway	Iron uptake and transport	57	0.278	12	0.048
R-HSA-2161522	pathway	Abacavir ADME	131	0.639	70	0.282
R-HSA-199220	pathway	Vitamin B5 metabolism	53	0.259	9	0.036
R-HSA-5205647	pathway	Mitophagy	115	0.561	71	0.286
R-HSA-9748784	pathway	Drug ADME	187	0.912	162	0.653
R-HSA-9749641	pathway	Aspirin ADME	178	0.868	154	0.621
R-HSA-4655427	pathway	SUMOylation of DNA methylation proteins	125	0.610	79	0.319
R-HSA-204005	pathway	COPII-mediated vesicle transport	54	0.263	14	0.056
R-HSA-9013026	pathway	RHOB GTPase cycle	62	0.302	19	0.077
R-HSA-9013106	pathway	RHOC GTPase cycle	64	0.312	19	0.077
R-HSA-6783310	pathway	Fanconi Anemia Pathway	150	0.732	122	0.492
R-HSA-6807070	pathway	PTEN Regulation	141	0.688	96	0.387

DIFLD+ / DIFLD- = total count of DIFLD-positive or -negative compounds containing the given substructure; *Ratio in+ / Ratio in-* = fraction of compounds within the respective group containing the given substructure.

Regarding the FAERS+DILIst dataset-based model, which used all three fingerprint-types, 26 pathways and four targets were among the 30 fingerprints (**Table 10**) with the highest scaled importance. Unsurprisingly, since PubChem BV performed notably worse, the best ranked BV appeared only at position 56 with a scaled importance of 0.072. Several pathways in relation to the pharmacokinetic, metabolism and transmembrane transport of certain drugs could be found within the top 20 fingerprints. Concerning the four most important targets, albumin, CYP2C8 and two members of the ATP-binding cassette (ABC) family, ABCG2 and ABCB1, were included.

Table 11. Top 30 most important features of the FAERS+DILIst US_GB model (targets)

Feature ID	Feature type	Description	DIFLD+	Ratio in+	DIFLD-	Ratio in-
P08183	target	ATP-dependent translocase ABCB1	116	0.57	55	0.22
P02768	target	Albumin	94	0.46	47	0.19
Q9UNQ0	target	ATP-binding cassette transporter ABCG2	53	0.26	9	0.04
P11712	target	CYP2C9	89	0.43	56	0.23
P08684	target	CYP3A4	149	0.73	122	0.49
P10635	target	CYP2D6	75	0.37	88	0.35
P10632	target	CYP2C8	70	0.34	24	0.10

P20815	target	CYP3A5	70	0.34	46	0.19
P24462	target	CYP3A7	55	0.27	59	0.24
P33261	target	CYP2C19	74	0.36	67	0.27
O95342	target	Bile salt export pump (BSEP)	42	0.20	16	0.06
P35348	target	Alpha-1A adrenergic receptor	37	0.18	58	0.23
Q9Y6L6	target	Organic anion transporter (OATP1B1)	47	0.23	13	0.05
P22309	target	UDP-glucuronosyltransferase 1A1	38	0.19	10	0.04
P02545	target	Prelamin-A/C	47	0.23	56	0.23
P05177	target	CYP1A2	54	0.26	48	0.19
P07550	target	Beta-2 adrenergic receptor	13	0.06	27	0.11
P35368	target	Alpha-1B adrenergic receptor	14	0.07	24	0.10
Q99720	target	Sigma non-opioid intracellular receptor 1	15	0.07	26	0.10
Q16236	target	Nuclear factor erythroid 2-related factor 2 (NRF2)	20	0.10	17	0.07
P03372	target	Estrogen receptor (ER-alpha)	21	0.10	25	0.10
P20813	target	CYP2B6	42	0.20	23	0.09
Q16665	target	Hypoxia-inducible factor 1-alpha	8	0.04	18	0.07
P21397	target	Amine oxidase	7	0.03	15	0.06
O75469	target	Pregnane X receptor (PXR)	38	0.19	25	0.10
P02763	target	Alpha-1-acid-glycoprotein 1	27	0.13	21	0.08
P16662	target	UDP-glucuronosyltransferase 2B7	27	0.13	10	0.04
Q14994	target	Constitutive androstane receptor (CAR)	24	0.12	9	0.04
Q92887	target	ATP-binding cassette sub family C member 2 ABCC2	32	0.16	9	0.04
O15245	target	Organic cation transporter 1 (OCT1)	18	0.09	19	0.08

DIFLD+ / *DIFLD-* = total count of DIFLD-positive or -negative compounds containing the given substructure; *Ratio in+* / *Ratio in-* = fraction of compounds within the respective group containing the given substructure.

Multiple CYP-enzymes were also present in the top 30 feature importance of the best performing FAERS+DILIst dataset-based model (**Table 11**) which only used the target fingerprint-type. Moreover, the three most important were ABCB1, albumin and ABCG2. Other features ranked within the top 30 included different adrenergic receptor subtypes, transporter proteins, UDP-glucuronosyltransferases (UGT), Prelamin A/C and some nuclear receptors such as CAR, PXR and ER, which are part of several AOP networks (AbdulHameed et al., 2019; Ortega-Vallbona et al., 2024).

Table 12. Top 30 most important features of the DILIst US_GB model (targets and pathways)

Feature ID	Feature type	Description	DIFLD+	Ratio in+	DIFLD-	Ratio in-
R-HSA-200425	pathway	Carnitine shuttle	71	0.133	8	0.027
R-HSA-9663891	pathway	Selective autophagy	379	0.711	150	0.512
P20815	target	CYP3A5	117	0.22	60	0.205
P10635	target	CYP2D6	183	0.343	104	0.355
P33261	target	CYP2C19	185	0.347	79	0.27
R-HSA-9753281	pathway	Paracetamol ADME	187	0.351	60	0.205
R-HSA-9837999	pathway	Mitochondrial protein degradation	107	0.201	29	0.099
R-HSA-9013406	pathway	RHOQ GTPase cycle	267	0.501	132	0.451
P07550	target	Beta-2 adrenergic receptor	23	0.043	30	0.102
R-HSA-9609507	pathway	Protein localization	244	0.458	81	0.276
R-HSA-879415	pathway	Advanced glycosylation end-product receptor signaling	263	0.493	102	0.348
R-HSA-202427	pathway	Phosphorylation of CD3 and TCR zeta chains	71	0.133	11	0.038
R-HSA-5579029	pathway	Metabolic disorders of biological oxidation enzymes	119	0.223	42	0.143
R-HSA-156581	pathway	Methylation	161	0.302	51	0.174
R-HSA-211999	pathway	CYP2E1 reactions	310	0.582	119	0.406
R-HSA-196849	pathway	Metabolism of water-soluble vitamins and cofactors	92	0.173	21	0.072
R-HSA-211945	pathway	Phase I - Functionalization of compounds	380	0.713	160	0.546
R-HSA-2142670	pathway	Synthesis of epoxy (EET) and dihydroxyeicosatrienoic acids (DHET)	332	0.623	131	0.447
R-HSA-418594	pathway	G alpha(i) signalling events	63	0.118	38	0.13
R-HSA-1632852	pathway	Macroautophagy	300	0.563	112	0.382
R-HSA-211859	pathway	Biological oxidations	344	0.645	144	0.491
R-HSA-5362517	pathway	Signaling by Retinoic Acid	179	0.336	58	0.198
R-HSA-9748784	pathway	Drug ADME	427	0.801	198	0.676
R-HSA-4551638	pathway	SUMOylation of chromatin organization proteins	317	0.595	141	0.481
R-HSA-5423646	pathway	Aflatoxin activation and detoxification	367	0.689	165	0.563
R-HSA-162582	pathway	Signal Transduction	332	0.623	159	0.543
R-HSA-109704	pathway	PI3K Cascade	108	0.203	27	0.092
R-HSA-168256	pathway	Immune System	288	0.54	136	0.464

R-HSA-199991	pathway	Membrane Trafficking	326	0.612	147	0.502
R-HSA-159418	pathway	Recycling of bile acids and salts	288	0.54	110	0.375

DIFLD+ / DIFLD- = total count of DIFLD-positive or -negative compounds containing the given substructure; *Ratio in+ / Ratio in-* = fraction of compounds within the respective group containing the given substructure.

To assess whether relevant differences or similarities exist, the feature importance was also calculated for the general DILI prediction models which employed the DILIst dataset (Thakkar et al., 2020). This was done for the DILI model that combined target and pathway fingerprints (**Table 12**) and for the DILI model that utilized targets only (**Table 13**). As with the FAERS+DILIst dataset-based model which used all fingerprint-types, 26 pathways and four targets were within the top 30 highest ranked features. Moreover, there were also some pathways related to the pharmacokinetics of drugs and metabolism of vitamins. Regarding the four targets, three different CYP-enzymes and the beta-2 adrenergic receptor were among the top 30 most important fingerprints.

Table 13. Top 30 most important features of the DILIst SMOTE_RF model (targets)

Feature ID	Feature type	Description	DIFLD+	Ratio in+	DIFLD-	Ratio in-
P11712	target	CYP2C9	210	0.394	69	0.235
P05177	target	CYP1A2	156	0.293	51	0.174
Q5BJF2	target	Sigma intracellular receptor	3	0.006	5	0.017
P01911	target	HLA class II histocompatibility antigen	21	0.039	0	0.000
P37231	target	Peroxisome proliferator-activated receptor gamma (PPAR-gamma)	36	0.068	4	0.014
Q9UNQ0	target	ATP-binding cassette transporter ABCG2	71	0.133	15	0.051
O94956	target	Organic anion transporter B (OATP-B)	35	0.066	5	0.017
O95342	target	Bile salt export pump (BSEP)	82	0.154	22	0.075
P10632	target	CYP2C8	110	0.206	32	0.109
P02768	target	Albumin	169	0.317	62	0.212
P08684	target	CYP3A4	314	0.589	146	0.498
P28482	target	MAP kinase 1	42	0.079	11	0.038
Q9NPD5	target	Organic anion transporter (OATP1B3)	45	0.084	9	0.031
P02545	target	Prelamin-A/C	136	0.255	62	0.212
Q99720	target	Sigma non-opioid intracellular receptor 1	38	0.071	28	0.096

O60656	target	UDP-glucuronosyltransferase 1A9	43	0.081	8	0.027
P11245	target	Arylamine N-acetyltransferase 2	14	0.026	1	0.003
P10275	target	Androgen receptor (AR)	70	0.131	23	0.078
Q9Y6L6	target	Organic anion transporter (OATP-C)	79	0.148	23	0.078
P03372	target	Estrogen receptor (ER-alpha)	77	0.144	32	0.109
P07550	target	Beta-2 adrenergic receptor	23	0.043	30	0.102
P04626	target	Receptor tyrosine-protein kinase erbB-2	13	0.024	0	0.000
P00352	target	Aldehyde dehydrogenase 1A1	33	0.062	10	0.034
P05181	target	CYP2E1	54	0.101	18	0.061
P04637	target	Cellular tumor antigen p53	46	0.086	17	0.058
Q96FL8	target	Multidrug and toxin extrusion protein 1 (MATE1)	31	0.058	8	0.027
P33261	target	CYP2C19	185	0.347	79	0.270
P11509	target	CYP2A6	36	0.068	9	0.031
P00918	target	Carbonic anhydrase 2	25	0.047	5	0.017
P10635	target	CYP2D6	183	0.343	104	0.355

DIFLD+ / DIFLD- = total count of DIFLD-positive or -negative compounds containing the given substructure; *Ratio in+ / Ratio in-* = fraction of compounds within the respective group containing the given substructure.

The 30 most important features of the only target-based DILI prediction model (**Table 13**) included several CYP-enzymes, OATPs, BSEP, albumin, prelamin-A/C, ABCG2 and two nuclear receptors (PPARγ, ER).

3.5 A visualization of the feature space

For the two models, which were ultimately employed for classification of the external compounds of interest, traditional t-SNE was used to visualize the fingerprint of each compound on a two-dimensional scatter plot (**Figure 9-10**). Moreover, the classification outcome of each drug was illustrated by color: positives in red or purple (TP = red, FN = purple) and negatives in green or blue (TN = green, FP = blue). Finally, the seven compounds of interest without an assigned DIFLD-label were highlighted in yellow.

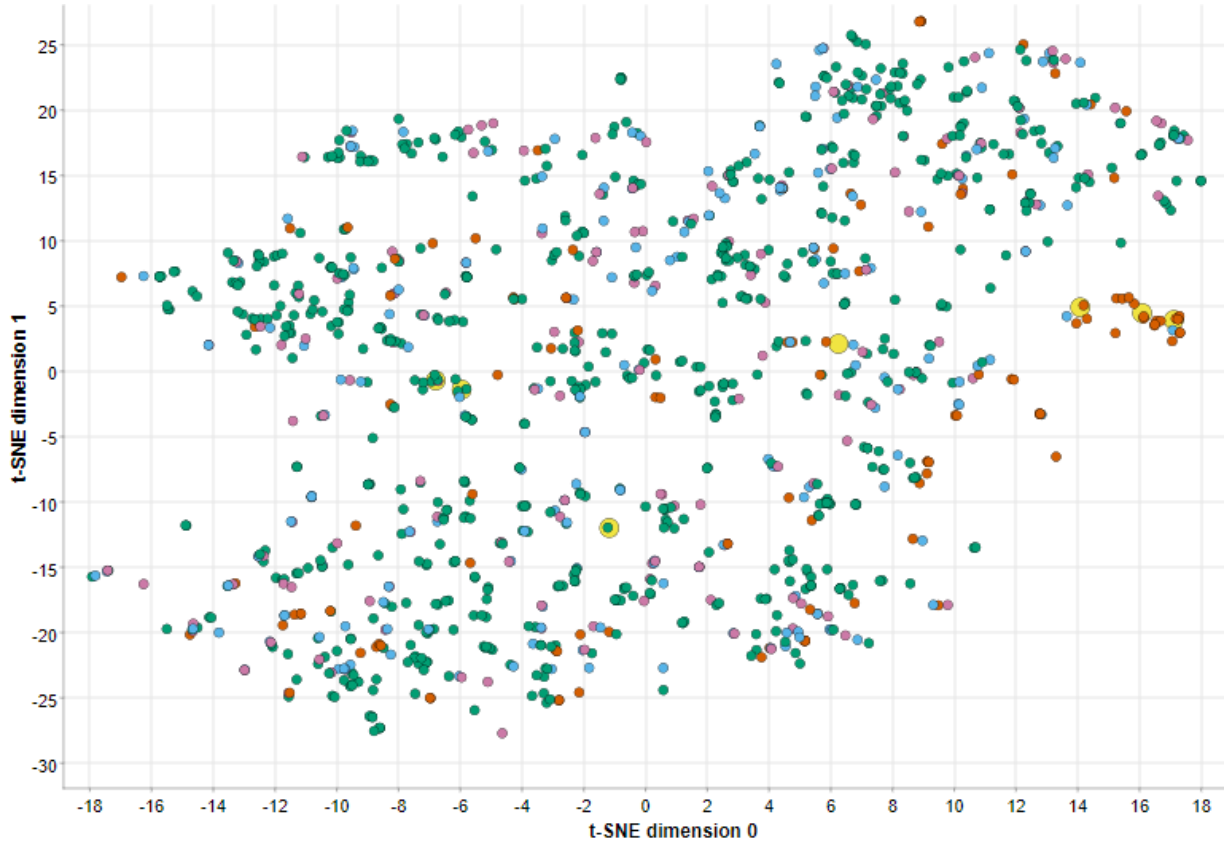


Figure 9. Screenshot of the interactive feature-space scatter plot of LJMU compounds (PubChem BV fingerprint) generated in KNIME. Prediction outcomes are color-coded: positives in red or purple (TP = red, FN = purple) and negatives in green or blue (TN = green, FP = blue). The seven external compounds of interest are highlighted in yellow.

Regarding the LJMU dataset, only the substructural PubChem BV fingerprint was employed to compute the two t-SNE dimensions. When examining the resulting scatter plot (**Figure 9**), several groupings mainly associated with DIFLD-negative compounds (green or blue) can be seen. However, for the DIFLD-positive compounds (red or purple) only one rather distinct cluster exists. Interestingly, three out of the seven external compounds of interest (yellow) are part of this cluster, namely propiconazole, tebuconazole and prochloraz.

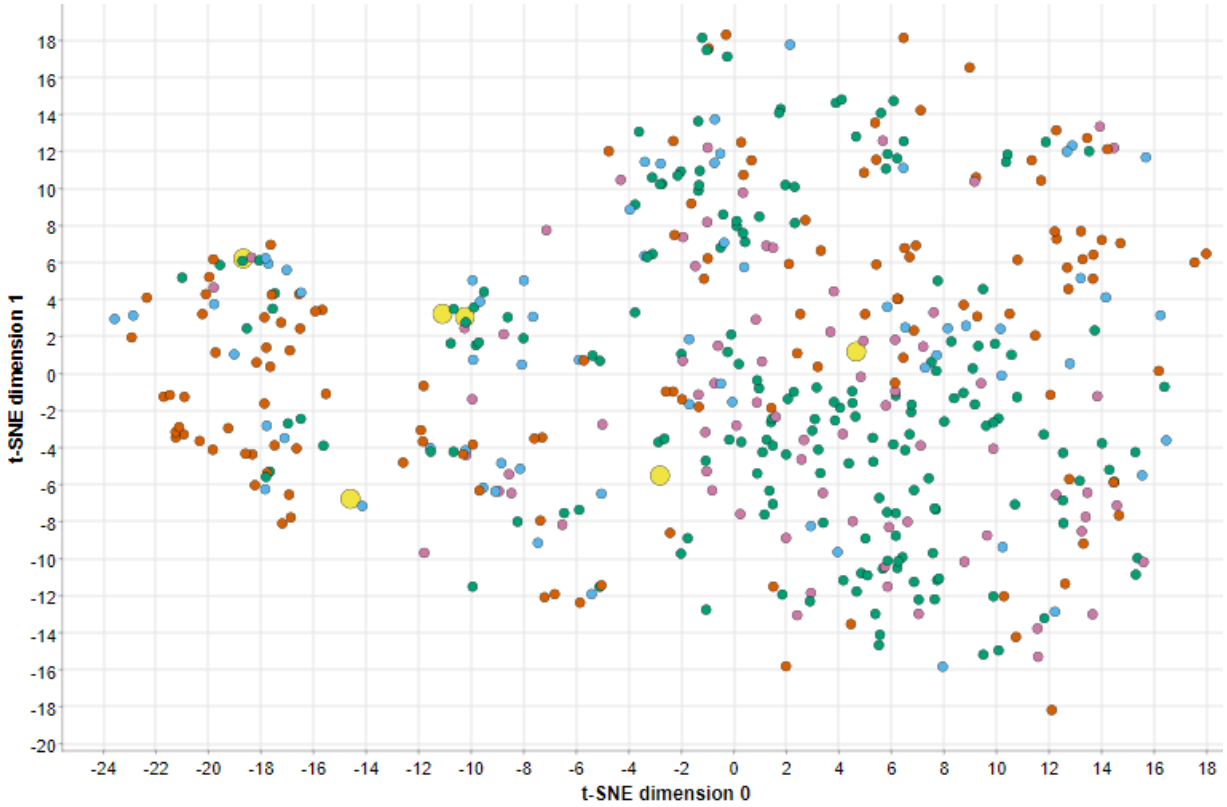


Figure 10. Screenshot of the interactive feature-space scatter plot of FAERS+DILIst compounds (combined fingerprint) generated in KNIME. Prediction outcomes are color-coded: positives in red or purple (TP = red, FN = purple) and negatives in green or blue (TN = green, FP = blue). The seven external compounds of interest are highlighted in yellow.

Some groupings were also evident regarding the FAERS+DILIst dataset-based model (**Figure 10**) which utilized all three fingerprint-types (PubChem BV, targets, pathways). However, in contrast to the scatter plot of the LJMU dataset (**Figure 9**), a greater number of clusters existed which were rather associated with true positive compounds (red). Moreover, near the true negative drugs (green), several false negatives (purple) were present. Concerning the external compounds of interest (yellow), no clear association with distinct groupings could be observed.

3.6 A visualization of the chemical space

The MolCompass tool (Sosnin, 2024) which computes ECFPs based on each compounds SMILES-code and ultimately establishes two coordinates for each drug employing deterministic t-SNE, was used for visualization of the chemical space. This approach enabled direct comparison of the ECFP-based chemical space between both datasets that were ultimately utilized for classification of the external compounds of interest. As with the

visualization of the feature space, the prediction outcome class for each chemical was illustrated by color and the test compounds of interest were highlighted in yellow.

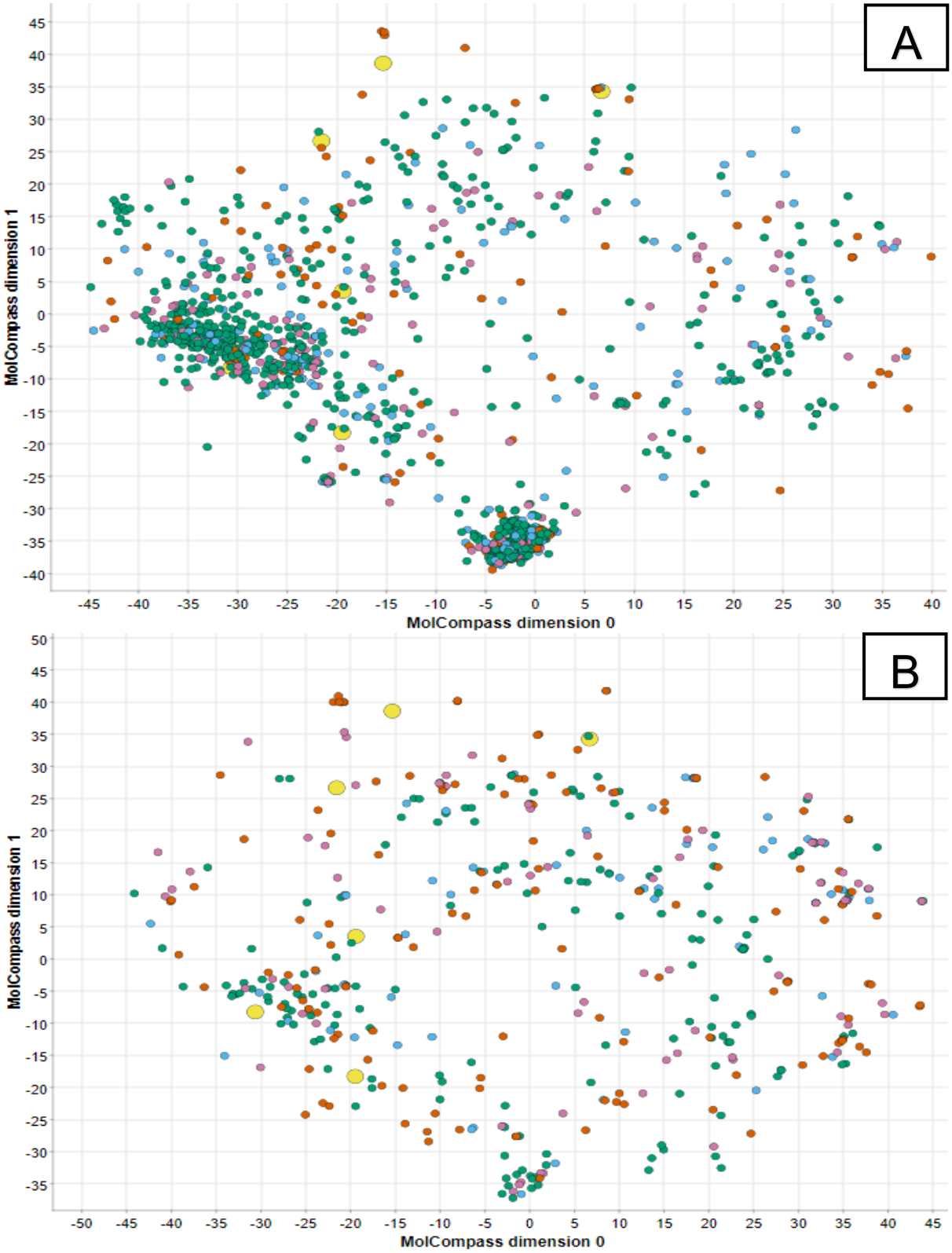


Figure 11. Comparison of the chemical-space scatter plots of the LJMU (**A**) and FAERS+DILIst (**B**) datasets. Prediction outcomes are color-coded: positives in red or purple (TP = red, FN = purple) and

negatives in green or blue (TN = green, FP = blue). The seven external compounds of interest are highlighted in yellow.

Interestingly, when comparing the resulting scatter plots of the LJMU (**Figure 11A**) and FAERS+DILIst (**Figure 11B**) datasets, two more densely populated areas are present in both chemical maps, located approximately in the same regions with respect to the coordinates. Moreover, for the LJMU dataset, these areas contain a remarkably high number of compounds, including many negatives (green or blue). Regarding groupings that are distinctly associated with one class, some smaller clusters with negative compounds can be found in the scatter plot of the LJMU dataset, with the largest one located around the -42/16 coordinate. However, concerning positive compounds (red or purple) no obvious groupings can be specified. In comparison, regarding the FAERS+DILIst dataset, only some smaller aggregations associated with positive or negative drugs are present. Unlike in the feature visualization of the LJMU dataset (**Figure 9**), which is determined by PubChem BVs, the ECFP-based chemical map (**Figure 11A**) does not arrange the external test molecules propiconazole, tebuconazole and prochloraz into a distinct cluster. Unsurprisingly, since MolCompass leads to a deterministic mapping of chemicals, all external compounds of interest (yellow) are located at the exact same coordinates, when comparing the chemical visualizations of the FAERS+DILIst (**Figure 11B**) and LJMU (**Figure 11A**) datasets. Finally, it is important to note that these chemical maps are just an illustration of the ECFPs and therefore do not reflect how the models actually perceived the compounds.

3.8 Prediction Results, Uncertainty and Applicability

Two models were selected to see how the seven test compounds of interest would be classified by them. While the chosen model which was trained on the LJMU dataset employed SMOTE_GB and only the substructural PubChem BV fingerprint, the selected FAERS+DILIst-based model applied SMOTE_RF and utilized all fingerprint-types, including target and pathway data. For the LJMU dataset, it was necessary to exclude all test compounds, since they were part of the original dataset. Regarding the FAERS+DILIst-based model, chlormequat chloride could not be predicted, as no relevant liver-expressed target could be identified by the designated workflow.

Two probability scores were calculated for each compound, which were then used to compute the prediction UC. A threshold of 0.35 was chosen to label molecules as either *"UC_Risk"* (≥ 0.35) or *"UC_OK"* (< 0.35).

The DtNN regarding test compounds and training compounds was employed to determine the model's AD, with a cut-off individually defined by the mean and SD of the DtNN between training molecules. This resulted in a threshold of 0.321 for the LJMU dataset-based model and a threshold of 0.516 for the FAERS+DILIst-based model, ultimately labeling each test compound as *"in_domain"* (< cut-off) or *"outside_domain"* (≥ cut-off). Moreover, this DtNN-cut-off was also applied to each training molecule, in order to examine how many training compounds the calculated threshold would define as not within the model's range. This resulted in 316 (26.89%) and 137 (30.31%) training molecules that would be classified as outside of domain, regarding both datasets, LJMU and FAERS+DILIst, respectively.

Table 14. Prediction outcomes, uncertainty and applicability domain of the external compounds of interest (LJMU dataset BV fingerprint SMOTE_GB model)

Compound ID	Compound name	DIFLD	Pred.	P (1)	P (0)	UC	UC flag	DtNN	AD flag
CHEMBL560579	propiconazole	?	1	0.989	0.011	0.011	UC_OK	0.136	In_domain
CHEMBL487186	tebuconazole	?	1	0.995	0.005	0.005	UC_OK	0.051	In_domain
CHEMBL118539	chlormequat chloride	0	0	0.177	0.823	0.177	UC_OK	0.192	in_domain
CHEMBL521027	cyprodinil	0	0	0.372	0.628	0.372	UC_Risk	0.056	in_domain
CHEMBL230001	azoxystrobin	?	0	0.075	0.925	0.075	UC_OK	0.197	in_domain
CHEMBL70971	carbendazim	?	0	0.067	0.933	0.067	UC_OK	0.076	in_domain
CHEMBL522782	prochloraz	?	1	0.957	0.043	0.043	UC_OK	0.166	in_domain

Based on:

Dataset	Fingerprint	Model	BA	MCC	Sensitivity	Specificity
LJMU†	BV	SMOTE_GB	0.610	0.213	0.427	0.793

†: Test compounds of interest were excluded.

Pred. = prediction made by the model; *P(1) / P(0)* = predicted probability of DIFLD-positivity (1) or negativity (0); *UC* = calculated uncertainty score; *UC flag* = uncertainty threshold flag; *DtNN* = Tanimoto distance to the nearest neighbor; *AD flag* = applicability domain flag.

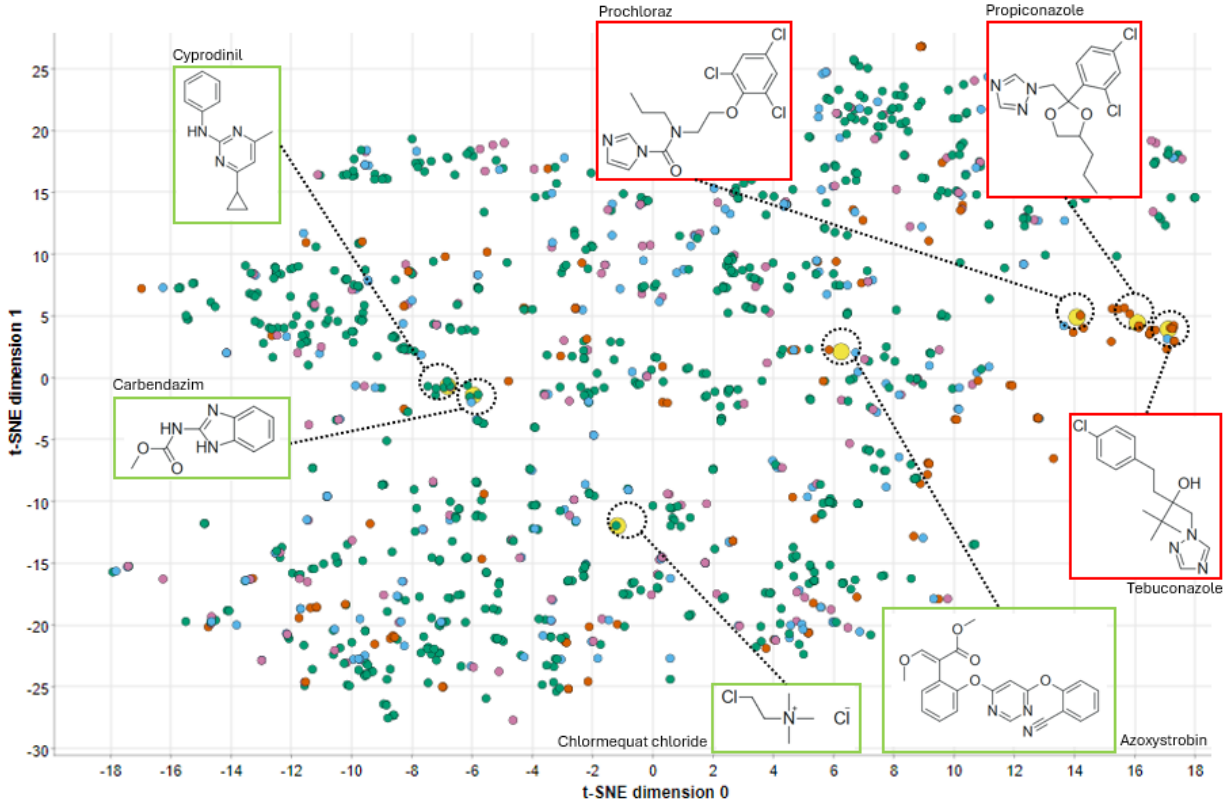


Figure 12. Feature-space scatter plot of LJMU compounds (PubChem BV). Prediction outcomes are color-coded: positives in red or purple (TP = red, FN = purple) and negatives in green or blue (TN = green, FP = blue). The seven external compounds of interest are highlighted in yellow, with their molecular structures displayed. Each structure is outlined with a colored frame indicating the prediction outcome (red = positive, green = negative).

Unsurprisingly, propiconazole, tebuconazole and prochloraz, the three compounds which were part of the distinct cluster of positives regarding the LJMU-dataset based model (**Figure 12**), were predicted as being steatogenic. Notably, the calculated UCs were especially low for these molecules (**Table 14**). Prochloraz, which was the most distant compound within the grouping, showed slightly higher UC. The remaining four test molecules were all classified as DIFLD-negative, two of them with a considerably higher UC. Since cyprodinil exceeded the threshold of 0.35, it was flagged as a prediction with "*UC_Risk*". Regarding the applicability domain, all seven compounds exhibited a DtNN clearly lower than the cut-off (0.321), classifying them as "*in_domain*". Interestingly, cyprodinil had the second lowest DtNN, despite displaying the highest UC.

Table 15. Prediction outcomes, uncertainty and applicability domain of the external compounds of interest (FAERS+DILIst dataset ALL fingerprint SMOTE_RF model)

Compound ID	Compound name	DIFLD	Pred.	P (1)	P (0)	UC	UC flag	DtNN	AD flag
CHEMBL560579	propiconazole	?	0	0.404	0.596	0.404	UC_Risk	0.359	in_domain
CHEMBL487186	tebuconazole	?	0	0.269	0.731	0.269	UC_OK	0.366	in_domain
CHEMBL118539	chlormequat chloride†	0	?	?	?	?	?	?	?
CHEMBL521027	cyprodinil	0	0	0.182	0.818	0.182	UC_OK	0.398	in_domain
CHEMBL230001	azoxystrobin	?	0	0.386	0.614	0.386	UC_Risk	0.404	in_domain
CHEMBL70971	carbendazim	?	0	0.137	0.863	0.137	UC_OK	0.260	in_domain
CHEMBL522782	prochloraz	?	0	0.310	0.690	0.310	UC_OK	0.501	in_domain

Based on:

Dataset	Fingerprint	Model	BA	MCC	Sensitivity	Specificity
FAERS+DILIst	ALL	SMOTE_RF	0.686	0.374	0.634	0.738

†: Chlormequat chloride was not predicted because no active target could be identified.

Pred. = prediction made by the model; *P(1) / P(0)* = predicted probability of DIFLD-positivity (1) or negativity (0); *UC* = calculated uncertainty score; *UC flag* = uncertainty threshold flag; *DtNN* = Tanimoto distance to the nearest neighbor; *AD flag* = applicability domain flag.

Examining the results of the FAERS+DILIst-based model (**Table 15**), the UCs were notably higher regarding every compound other than cyprodinil. Moreover, the predictions for propiconazole, tebuconazole and prochloraz were contradictory when compared with those from the LJMU dataset-based model. However, their UC scores were considerably higher, with the DIFLD-negative classification of tebuconazole being flagged as "*UC_Risk*". Moreover, the prediction of azoxystrobin also surpassed the UC threshold, labeling its classification as uncertain. Regarding the applicability domain, all test compounds DtNN were clearly higher in this dataset. However, since this was also true for the DtNN values between training compounds, the calculated cut-off which defines whether a molecule is "*in_domain*", or "*outside_domain*" was larger compared to that of the LJMU dataset. Hence, regarding the FAERS+DILIst-based model, all external test compounds were still labeled as within domain. Finally, the molecule with the lowest UC, carbendazim, also exhibited the lowest DtNN.

3.9 Uncertainty filtering

To see how filtering out uncertain compounds would influence the results of the ten-fold cross validation, the 0.35 cut-off was also applied to the determined UC scores of each molecule in the LJMU and FAERS+DILIst dataset.

Table 16. Impact of uncertainty filtering on cross-validation results of the LJMU SMOTE_GB model (PubChem BV)

UC filtering	Dataset	FP	Model	BA	MCC	Sensitivity	Specificity
UC_OK	LJMU†	BV	SMOTE_GB	0.621	0.245	0.414	0.827
Unfiltered	LJMU†	BV	SMOTE_GB	0.610	0.213	0.427	0.793
UC_Risk	LJMU†	BV	SMOTE_GB	0.552	0.088	0.500	0.604

†: External test compounds of interest were excluded.

UC = uncertainty; *FP* = fingerprint-type; *BA* = balanced accuracy; *MCC* = Matthews correlation coefficient; *BV* = PubChem Bitvectors (substructural fingerprint); *SMOTE* = synthetic minority oversampling technique.

997 (84.78%) of the LJMU compounds had a prediction uncertainty of under 0.35. When filtering for these molecules before statistical analysis, all performance scores except for the sensitivity went up (**Table 16**). Moreover, when only considering the 179 (15.22%) compounds classified as uncertain (≥ 0.35), the performance was lower, with significant decreases observed in BA, MCC, specificity, accuracy and precision. However, the sensitivity reached its highest value for this subset, showing the same trend as before.

Table 17. Impact of uncertainty filtering on cross-validation results of the FAERS+DILIst SMOTE_RF model (combined fingerprint)

UC_filtering	Dataset	FP	Model	BA	MCC	Sensitivity	Specificity
UC_OK	FAERS+DILIst	ALL	SMOTE_RF	0.747	0.498	0.681	0.813
Unfiltered	FAERS+DILIst	ALL	SMOTE_RF	0.686	0.374	0.634	0.738
UC_Risk	FAERS+DILIst	ALL	SMOTE_RF	0.588	0.175	0.573	0.602

UC = uncertainty; *FP* = fingerprint-type; *BA* = balanced accuracy; *MCC* = Matthews correlation coefficient; *SMOTE* = synthetic minority oversampling technique.

Regarding the FAERS+DILIst dataset, 276 (60.93%) compounds exhibited a prediction UC score of under 0.35. Unlike the other dataset (**Table 16**), filtering for these molecules before performance evaluation led to higher results across all metrics, including sensitivity (**Table 17**). This trend could also be examined when restricting this dataset to the 177 (39.07%) compounds classified as uncertain (≥ 0.35), which resulted in lower performances.

4 Discussion

The main purpose of this thesis was to employ the KNIME-workflow from Helmke & Ecker, which is based on the approach by Füzi et al., to generate binary DIFLD prediction models from substructural and biological fingerprints, and to assess their predictive performance. Of particular interest was the feature importance computed by these models, which may allow for potential hypothesis formulation. Ultimately, seven compounds of interest were chosen to assess their classification by the best-performing models.

To summarize results, it was found that for many non-pharmaceutical compounds, the workflow failed to retrieve sufficient target and pathway information. However, the best-performing model, regarding the FAERS+DILIst dataset, which mainly consisted of approved drugs, was target-based and achieved a sensitivity of 0.79 and an MCC of 0.42. Across all pharmaceutical-based datasets, CYP-enzymes, ABC-transporters, prelamin A/C, albumin and certain nuclear receptors were frequently among the top 30 most important ranked features. The FAERS+DILIst based model which used all three fingerprint types and the LJMU dataset-based model which only employed substructural PubChem BV were chosen for classification of the external test compounds of interest. Predictions for three out of seven test molecules were contradictory, with the FAERS+DILIst based model exhibiting substantially higher uncertainties regarding these compounds.

4.1 Datasets

Given that the designated workflow did not manage to retrieve sufficient biological target data for the originally intended binary DIFLD dataset (LJMU) to encode the compounds for biological fingerprint-based modeling, a limitation of the applied approach became evident. The workflow appears to retrieve significantly more relevant data for drugs that are approved for human use. Possible reasons could be a general lack of data or the fact that the workflow, beyond PubChem and ChEMBL, incorporates other databases, such as DrugBank and PharmGKB with a specific focus on pharmaceuticals (Knox et al., 2024; Whirl-Carrillo et al., 2021). This likely explains why, for compounds from datasets which are based on post-marketing reports or focused on pharmaceuticals in general, sufficient target-information can usually be obtained to encode them for modeling, as seen here for the FAERS+DILIst dataset. Helmke & Ecker who utilized this information retrieval workflow to generate a biological fingerprint-based cholestasis prediction model, employed a

dataset that was also derived from adverse event reports and consisted mainly of human-used pharmaceuticals (Kotsampasakou & Ecker, 2017).

When comparing the DILI-positive class from DILIst (Thakkar et al., 2020) with DIFLD-positives from the curation of FAERS (**Figure 7**), 140 (68%) of the positive labels were concordant and 35 (17%) of the possibly steatogenic drugs could not be found in the DILIst dataset, indicating reasonable validity of this self-curated dataset. Among the remaining 30 (15%) contradicting compounds, several frequently prescribed drugs were present, including opioids which can only be found in the DILI-negative class regarding DILIst. This may be due to the way FAERS was curated. Since an absolute cut-off of 50 cases was chosen to define DIFLD-positive compounds, regardless of prescription frequency, more commonly prescribed drugs are more likely to appear. Taking this into consideration, an alternative way of curating the data, which also accounts for the total number of case reports per drug, may provide greater accuracy. Interestingly, among the 35 (17%) DIFLD-positive compounds that could not be found in the DILIst dataset, a relatively high number of tyrosine kinase inhibitors were observed. Since several drugs of this class were also among the 140 (68%) concordant DIFLD-positive compounds, the positive assignment seems legitimate. Moreover, hepatotoxicity has been reported for many of these tyrosine kinase inhibitors (Paech et al., 2017; L. Zhu et al., 2024), including ponatinib and nilotinib, which were classified as steatogenic according to the curation of FAERS but are missing in the DILIst dataset. Since these two compounds were retrieved by the additional curation and filtering of individual case reports from the FAERS line listing table, this approach may be advantageous compared to curation based solely on the total number of reported cases.

4.2 Model Performances

Performance evaluation results regarding the LJMU dataset-based model (**Table 3**), which relied solely on PubChem BV, were highly comparable to those reported by Jain et al., who also created a steatosis prediction model. Although Jain et al. employed ToxPrint chemotypes as the substructural fingerprint, the sensitivity (0.47) and specificity (0.78) of both structure-based models were nearly identical. This similarity is likely due to the fact

that the LJMU dataset largely consisted of compounds from Jain et al., resulting in a substantial overlap.

While the FAERS+DILIst-based model which only employed the target fingerprint performed well with a sensitivity of 0.79 and an MCC of 0.42 (**Table 4**), the observed discrepancy in active targets per drug between the DIFLD-positives and negatives may indicate an unintended dataset bias. Thus, it may be possible that the model's classification of compounds into DIFLD-positive or -negative is influenced by this bias rather than by biological signals associated with steatosis. Moreover, the discrepancy in actives was notably smaller within the poorest performing dataset DILIst further supporting this assumption. When filtering results from the ten-fold cross-validation for positive drugs with fewer than the median number of actives per compound and negative drugs with more than the median number of actives per compound, the BA decreased to 0.5, indicating that these drugs were frequently misclassified. However, given the fact that the designated workflow filters for targets with high expression in liver tissue, a higher number of interactions regarding DIFLD- or DILI-positive compounds seems plausible and hence may explain the observed discrepancy in active-count between the classes. Moreover, since some steatosis-relevant targets identified in the AOPs and literature were ranked among the 30 most important features, this may suggest that the model classifies not only by the sheer number of highly liver-expressed active targets, but also by particularly considering targets directly related to DIFLD. This assumption may be further supported by the observation that, when filtering all compounds for those with an active-count greater than the median, several negative drugs were still correctly classified (**Table 8**), with a sensitivity of 0.88 and a specificity of 0.42. Since Helmke & Ecker generated a cholestasis-prediction model utilizing the same fingerprint retrieval workflow and did not report on the active-count per drug for each class (Helmke & Ecker, 2025), it is unclear whether this discrepancy also occurred in their model. In the future, users of the workflow should take this aspect into account and share their observations, particularly regarding its impact on model performance.

Füzi et al., the creators of the original Path4Drug KNIME workflow (Füzi et al., 2021), generated DILI prediction models also employing target and pathway information. Their best performing models' predictive accuracies (≈ 0.69) were higher than those of the

DILIst-based models in this thesis (≈ 0.57) (**Table 6**). However, this is likely because Füzi et al. utilized the DILIrank dataset (M. Chen et al., 2016), which comprises four categories – mostDILI, lessDILI, ambiguousDILI, and noDILI. While Füzi et all restricted their analysis to the highest-confidence labels (mostDILI and noDILI), the binary DILIst dataset used here augments DILIrank with additional compounds and data from other sources, resulting in a substantially larger total number of drugs (Thakkar et al., 2020). Given the lower overall confidence of DILIst, the weaker predictive performance observed in this thesis appears reasonable.

4.3 Feature importance

One major aim of this thesis was to provide the feature importance to enable comparison with literature and previous findings, and for the potential generation of new hypotheses. When comparing the top 30 most important fingerprints with the feature importance results of other studies that generated hepatotoxicity prediction models employing target or pathway information as fingerprints, several parallels were observed. Liu et al., who created a DILI classification model utilizing predicted protein targets as descriptors, provided a table with the 19 highest rated proteins, including CYP1A2 and CYP2C9. These targets were the top two most important proteins regarding the DILI prediction model of this thesis (**Table 13**) and also among the highest rated fingerprints of the FAERS+DILIst-based DIFLD prediction model (**Table 11**). CYP2C9 was also included in the top 10 descriptors of a hepatotoxicity prediction model from Füzi et al. Regarding pathways, two of the 10 most important fingerprints, the heme degradation and metabolism of porphyrins pathways from Reactome (Milacic et al., 2024) were also among the 10 highest scoring features of the FAERS+DILIst-based model (**Table 10**) which employed all three fingerprint types. Some parallels could also be found when comparing with feature importance results from the cholestasis prediction model generated by Helmke & Ecker. This involved the targets albumin and BSEP (**Table 11**), as well as the pathway sphingolipid de novo biosynthesis (**Table 10**). Interestingly, Helmke & Ecker noted that Jackson et al. had previously described dysregulation of this pathway as being associated with steatosis in patients with NAFLD.

Although other proteins scored higher in the fingerprint importance, several targets regarding MIEs from the AOP network for liver steatosis (Ortega-Vallbona et al., 2024) were among the 30 most important features, regarding the FAERS+DILIst-based DIFLD prediction model, which only employed the target-fingerprint (**Table 11**). These proteins were nuclear receptors, including NRF2, ER, PXR, and CAR. Interestingly, compared to higher scoring proteins, the difference in active-occurrence ratios between the DIFLD-positive and negative class regarding these targets were considerably low, with ER exhibiting a ratio of 0.10 in both classes. One possible explanation could be that the fingerprint does not distinguish between activation and inhibition of these nuclear receptors. It is important to note that a value of 1(active) in the feature matrix only indicates a binding event, without specifying whether activation or inhibition occurs. Ultimately, this may explain why these nuclear receptors, which are thought to be highly associated with DIFLD (AbdulHameed et al., 2019; López-Pascual et al., 2024; Ortega-Vallbona et al., 2024), scored lower in importance compared to certain other proteins. Moreover, PPARγ, which was the only nuclear receptor among the 30 most important features of the DILIst dataset-based model (**Table 13**), showed a relatively high difference in active occurrence ratios and was placed fifth in the ranking.

Interestingly, Schinderle et al. found that an accumulation of prelamin, which leads to nuclear morphology changes, can be observed in MASLD patients. Moreover, they report that the accumulation is due to downregulation of the prelamin-processing enzyme ZMPSTE24, which is inhibited by several antiretrovirals used in HIV-therapy (Schinderle et al., 2024). The morphologic changes of the nucleus result in different FOXA2 binding (Schinderle et al., 2024), a transcription factor important for liver function and development (Aghadi et al., 2022). Schinderle et al. analyzed the changed FOXA2 binding sites in patients with moderate to severe steatosis and, through Ingenuity Pathway Analysis, identified some nuclear receptors, including CAR. Additionally, they identified ER as a regulator of genes bound by FOXA2 and pathways like “*NRF2 Stress Response*”, “*PXR Activation*” and “*Fatty Acid Metabolism*” when analyzing differences in gene expression in MASLD-patients. Finally, regarding female patients with moderate steatosis, they also detected the upregulation of several nuclear receptors, such as FXR, CAR, LXR and PXR (Schinderle et al., 2024). Notably, NRF2, ER, PXR, and CAR – all the nuclear receptors ranked among the top features of the best-performing DIFLD prediction model in this

thesis (**Table 11**) – were also identified as being associated with hepatic steatosis by Schinderle et al. Moreover, prelamin, which interestingly showed the same actives-occurrence ratio between DIFLD-positive and -negative compounds, was also among the most important targets in the feature importance. However, ZMPSTE24, the prelamin-processing enzyme that is inhibited by certain antiretroviral drugs and as stated by Schinderle et al., is responsible for the subsequent accumulation of prelamin, was not detected.

4.4 A cluster of DIFLD-positives in the feature space

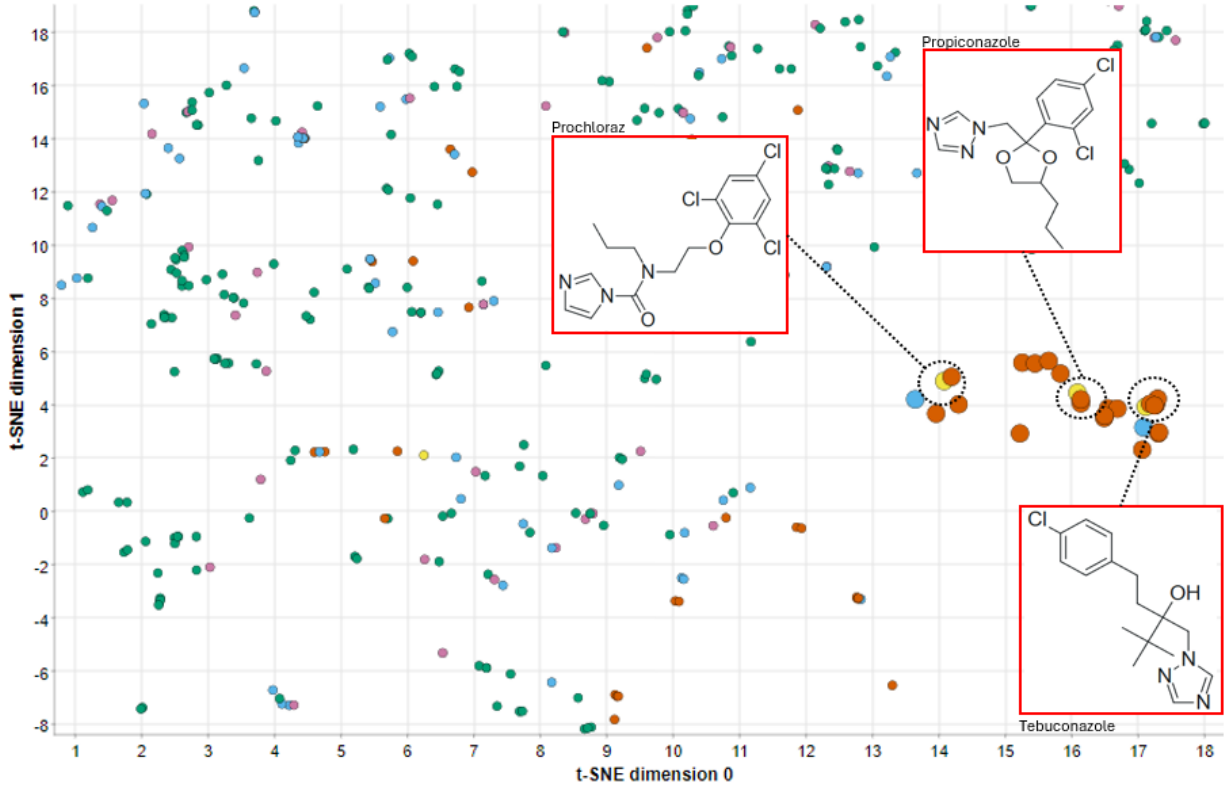


Figure 13. Zoomed-in excerpt of the feature-space scatter plot of LJMU compounds (PubChem BV). Prediction outcomes are color-coded: positives in red or purple (TP = red, FN = purple) and negatives in green or blue (TN = green, FP = blue). The seven external compounds of interest are highlighted in yellow, with their molecular structures displayed. Each structure is outlined with a colored frame indicating the prediction class (red = positive, green = negative).

Interestingly, as highlighted in **Figure 13**, a rather distinct cluster of positive compounds could be observed, with only two false positives (blue) out of 27 molecules. Most of these compounds were azoles, including mostly pesticides such as difenoconazole and only few azole-antifungal drugs approved for human use, such as ketoconazole. Although

ketoconazole was one of the two incorrectly predicted negatives (blue), as it was labeled DIFLD-negative in the LJMU dataset, it is notably classified as DILI-positive in the DILIst dataset (Thakkar et al., 2020) and even assigned to the most-DILI-concern category in the DILIrank dataset (M. Chen et al., 2016), which raises doubts regarding the validity of its classification in the LJMU dataset. Unsurprisingly, propiconazole, tebuconazole and prochloraz – three external test compounds of interest which were part of this DIFLD-positives cluster – were ultimately classified as being steatogenic, with remarkably low UCs (**Table 14**). Propiconazole and tebuconazole contain a triazole ring in their structure, meaning that, like difenoconazole, they carry BV496 (*N:N-C-[#1]*), which was ranked as the second most important substructural feature (**Table 9**). Moreover, prochloraz and propiconazole possess two or more chloro groups, indicating that they contain BV38 *(>= 2 Cl*), which was identified as the third most important feature. Although the best LJMU dataset-based model achieved only modest performance in the ten-fold cross-validation with a BA of 0.63 (**Table 3**), the prediction-outcome of these three compounds is further supported by their spatial association with the cluster of DIFLD-positives (**Figure 13**). Finally, their spatial proximity to this grouping also provides a rationale for their low UC scores.

4.4 Limitations of this thesis

This thesis has several limitations that should be acknowledged. Firstly, regarding the self-curated binary classification DIFLD datasets with positives derived from FAERS, only limited validity can be assumed, as these datasets were manually curated through self-defined thresholds. Moreover, the biological fingerprint matrix, which included targets and pathways, was based solely on actives (1) retrieved from the chosen databases. Consequently, inactives (0) merely indicated the absence of available information, rather than actually proven inactivity. This ultimately implies that compounds with little associated data in the respective databases exhibit substantially fewer actives, even though this does not necessarily reflect reality. Logically, models based on such biological fingerprints cannot be directly used for classification of novel molecules with little or no available biological data. Furthermore, it appeared that the models relying on biological fingerprints tended to classify compounds with many actives as DIFLD-positive and compounds with few actives as DIFLD-negative, possibly due to an observed discrepancy in active-count

between positive and negative training molecules. Considering these limitations, the presented classifications of the seven external test compounds of interest should be interpreted with caution, particularly regarding the FAER-DILIst-based model. Finally, when analyzing the feature importance for possible hypothesis generation, it is crucial to recognize that the computed importance of a descriptor is merely based on correlation and does not necessarily reflect any causal connections.

5 Conclusion and outlook

This thesis applied and adapted the cholestasis-based KNIME workflow from Helmke & Ecker to generate binary DIFLD prediction models, employing PubChem BV as structural features and interpretable biological fingerprints (targets and pathways). The feature importance was further presented to potentially enable hypothesis generation regarding pathological mechanisms of DIFLD. Ultimately, two of the best-performing models – one based solely on structural fingerprints and the other based on a combination of both PubChem BV and biological data – were chosen to see how they would classify seven compounds of interest.

The best-performing model based on the FAERS+DILIst dataset, which mainly comprised approved drugs, was target-driven and reached a sensitivity of 0.79 and an MCC of 0.42. In contrast, insufficient target and pathway information was retrieved for the LJMU dataset containing mostly non-pharmaceuticals. However, the best performing model for this dataset, which only employed a structure-based fingerprint, reached a sensitivity of 0.47 and an MCC of 0.24. Across the target-based models, nuclear receptors (NRF2, ER, PXR, CAR), albumin, prelamin A/C, CYP-enzymes, and ABC transporters were among the 30 top-ranked features. For three compounds of interest, the two chosen models made contradictory predictions; however, the LJMU dataset-based model showed notably lower UCs, and the feature space analysis revealed a close spatial proximity of these molecules to a rather distinct cluster of DIFLD-positives.

Although this thesis highlighted certain limitations in applying biological fingerprints obtained through the designated workflow for the generation of toxicity prediction models, reasonable performance was achieved within ten-fold cross-validation. Furthermore, the feature importance – especially of the target fingerprint – yielded directly interpretable and comparable results, which showed similarities to findings reported in the literature.

Finally, a combination of this data retrieval workflow with an automated curation of FAERS could provide a comprehensive pipeline to construct toxicity prediction models for any adverse events included in FAERS. Such models would rely on readily interpretable biological fingerprints and thus have the potential to deliver new insights into the underlying pathomechanisms. While data scarcity might present a challenge, especially when attempting to classify novel compounds, emerging data imputation methods based on neural networks (Cerimagic et al., 2025) may offer a promising strategy to close this gap.

References

AbdulHameed, M. D. M., Pannala, V. R., & Wallqvist, A. (2019). Mining Public Toxicogenomic Data Reveals Insights and Challenges in Delineating Liver Steatosis Adverse Outcome Pathways. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01007

Aghadi, M., Elgendy, R., & Abdelalim, E. M. (2022). Loss of FOXA2 induces ER stress and hepatic steatosis and alters developmental gene expression in human iPSC-derived hepatocytes. *Cell Death & Disease*, *13*(8), 713. https://doi.org/10.1038/s41419-022-05158-0

Allaway, R. J., La Rosa, S., Guinney, J., & Gosline, S. J. C. (2018). Probing the chemical–biological relationship space with the Drug Target Explorer. *Journal of Cheminformatics*, *10*(1), 41. https://doi.org/10.1186/s13321-018-0297-4

Allison, R., Guraka, A., Shawa, I. T., Tripathi, G., Moritz, W., & Kermanizadeh, A. (2023). Drug induced liver injury—A 2023 update. *Journal of Toxicology and Environmental Health. Part B, Critical Reviews*, *26*(8), 442–467. https://doi.org/10.1080/10937404.2023.2261848

Armstrong, L. E., & Guo, G. L. (2017). Role of FXR in Liver Inflammation during Nonalcoholic Steatohepatitis. *Current Pharmacology Reports*, *3*(2), 92–100. https://doi.org/10.1007/s40495-017-0085-2

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, *7*(1), 20. https://doi.org/10.1186/s13321-015-0069-3

Beisken, S., Meinl, T., Wiswedel, B., de Figueiredo, L. F., Berthold, M., & Steinbeck, C. (2013). KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinformatics*, *14*(1), 257. https://doi.org/10.1186/1471-2105-14-257

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2009). KNIME - the Konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, *11*(1), 26–31. https://doi.org/10.1145/1656274.1656280

Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, *15*(1), 73. https://doi.org/10.1186/s13321-023-00743-7

Borgne-Sanchez, A., & Fromenty, B. (2025). Mitochondrial dysfunction in drug-induced hepatic steatosis: Recent findings and current concept. *Clinics and Research in Hepatology and Gastroenterology*, *49*(3), 102529. https://doi.org/10.1016/j.clinre.2025.102529

Cerimagic, T., Sosnin, S., & Ecker, G. (2025). *A Multi-Task Learning Approach for Data Imputation of Compound Bioactivity Values for the SLC Transporter Superfamily*. ChemRxiv. https://doi.org/10.26434/chemrxiv-2025-7vtgv-v3

Cha, G.-W., Moon, H.-J., & Kim, Y.-C. (2021). Comparison of Random Forest and Gradient Boosting Machine Models for Predicting Demolition Waste Based on Small Datasets and Categorical Variables. *International Journal of Environmental Research and Public Health*, *18*(16), Article 16. https://doi.org/10.3390/ijerph18168530

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., & Cesareni, G. (2007). MINT: The Molecular INTeraction database. *Nucleic Acids Research*, *35*(suppl_1), D572–D574. https://doi.org/10.1093/nar/gkl950

Chen, C.-H., Tanaka, K., Kotera, M., & Funatsu, K. (2020). Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR

applications. *Journal of Cheminformatics*, *12*(1), 19. https://doi.org/10.1186/s13321-020-0417-9

Chen, M., Suzuki, A., Thakkar, S., Yu, K., Hu, C., & Tong, W. (2016). DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discovery Today*, *21*(4), 648–653. https://doi.org/10.1016/j.drudis.2016.02.015

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. https://doi.org/10.1186/s13040-023-00322-4

Clifford, B. L., Sedgeman, L. R., Williams, K. J., Morand, P., Cheng, A., Jarrett, K. E., Chan, A. P., Brearley-Sholto, M. C., Wahlström, A., Ashby, J. W., Barshop, W., Wohlschlegel, J., Calkin, A. C., Liu, Y., Thorell, A., Meikle, P. J., Drew, B. G., Mack, J. J., Marschall, H.-U., … de Aguiar Vallim, T. Q. (2021). FXR activation protects against NAFLD via bile-acid-dependent reductions in lipid absorption. *Cell Metabolism*, *33*(8), 1671-1684.e4. https://doi.org/10.1016/j.cmet.2021.06.012

Cotterill, J., Price, N., Rorije, E., & Peijnenburg, A. (2020). Development of a QSAR model to predict hepatic steatosis using freely available machine learning tools. *Food and Chemical Toxicology*, *142*, 111494. https://doi.org/10.1016/j.fct.2020.111494

del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., Perfetto, L., How, K., Ratan, P., Shirodkar, G., Lu, O., Mészáros, B., Watkins, X., Pundir, S., Licata, L., Iannuccelli, M., Pellegrini, M., Martin, M. J., Panni, S., … Hermjakob, H. (2022). The IntAct database: Efficient access to fine-grained molecular interaction data. *Nucleic Acids Research*, *50*(D1), D648–D653. https://doi.org/10.1093/nar/gkab1006

Dirven, H., Vist, G. E., Bandhakavi, S., Mehta, J., Fitch, S. E., Pound, P., Ram, R., Kincaid, B., Leenaars, C. H. C., Chen, M., Wright, R. A., & Tsaioun, K. (2021). Performance of preclinical models in predicting drug-induced liver injury in humans: A systematic review. *Scientific Reports*, *11*(1), 6403. https://doi.org/10.1038/s41598-021-85708-2

Duy, H. A., & Srisongkram, T. (2025). Protecting your skin: A highly accurate LSTM network integrating conjoint features for predicting chemical-induced skin irritation. *Journal of Cheminformatics*, *17*(1), 39. https://doi.org/10.1186/s13321-025-00980-y

Escher, S. E., Aguayo-Orozco, A., Benfenati, E., Bitsch, A., Braunbeck, T., Brotzmann, K., Bois, F., van der Burg, B., Castel, J., Exner, T., Gadaleta, D., Gardner, I., Goldmann, D., Hatley, O., Golbamaki, N., Graepel, R., Jennings, P., Limonciel, A., Long, A., … Fisher, C. (2022). Integrate mechanistic evidence from new approach methodologies (NAMs) into a read-across assessment to characterise trends in shared mode of action. *Toxicology in Vitro*, *79*, 105269. https://doi.org/10.1016/j.tiv.2021.105269

FDA Adverse Events Reporting System (FAERS) Public Dashboard—FDA Adverse Events Reporting System (FAERS) Public Dashboard | Arbeitsblatt—Qlik Sense. (n.d.). Retrieved August 4, 2025, from https://fis.fda.gov/sense/app/95239e26-e0be-42d9-a960-9a5f7f1c25ee/sheet/7a47a261-d58b-4203-a8aa-6d3021737452/state/analysis

Füzi, B., Gurinova, J., Hermjakob, H., Ecker, G. F., & Sheriff, R. (2021). Path4Drug: Data Science Workflow for Identification of Tissue-Specific Biological Pathways Modulated by Toxic Drugs. *Frontiers in Pharmacology*, *12*. https://doi.org/10.3389/fphar.2021.708296

Füzi, B., Mathai, N., Kirchmair, J., & Ecker, G. F. (2023). Toxicity prediction using target, interactome, and pathway profiles as descriptors. *Toxicology Letters*, *381*, 20–26. https://doi.org/10.1016/j.toxlet.2023.04.005

García-Cortés, M., Toro-Ortiz, J. P., & García-García, A. (2023). Deciphering the liver enigma: Distinguishing drug-induced liver injury and metabolic dysfunction-associated steatotic liver disease—a comprehensive narrative review. *Exploration of Digestive Diseases*, *2*(6), Article 6. https://doi.org/10.37349/edd.2023.00034

Hao, M., Wang, Y., & Bryant, S. H. (2014). An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analytica Chimica Acta*, *806*, 117–127. https://doi.org/10.1016/j.aca.2013.10.050

Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P. H., Davenport, A. P., Spedding, M., & Davies, J. A. (2024). The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Research*, *52*(D1), D1438–D1449. https://doi.org/10.1093/nar/gkad944

Helmke, P. S., & Ecker, G. F. (2025). Refining Drug-Induced Cholestasis Prediction: An Explainable Consensus Model Integrating Chemical and Biological Fingerprints. *Journal of Chemical Information and Modeling*, *65*(11), 5301–5316. https://doi.org/10.1021/acs.jcim.4c02363

Hosack, T., Damry, D., & Biswas, S. (2023). Drug-induced liver injury: A comprehensive review. *Therapeutic Advances in Gastroenterology*, *16*, 17562848231163410. https://doi.org/10.1177/17562848231163410

Index of /pub/databases/chembl/UniChem/data/wholeSourceMapping. (n.d.). Retrieved August 4, 2025, from https://ftp.ebi.ac.uk/pub/databases/chembl/UniChem/data/wholeSourceMapping/

Jain, S., Norinder, U., Escher, S. E., & Zdrazil, B. (2021). Combining In Vivo Data with In Silico Predictions for Modeling Hepatic Steatosis by Using Stratified Bagging and Conformal Prediction. *Chemical Research in Toxicology*, *34*(2), 656–668. https://doi.org/10.1021/acs.chemrestox.0c00511

Jin, H., Zhang, C., Zwahlen, M., von Feilitzen, K., Karlsson, M., Shi, M., Yuan, M., Song, X., Li, X., Yang, H., Turkez, H., Fagerberg, L., Uhlén, M., & Mardinoglu, A. (2023). Systematic transcriptional analysis of human cell lines for gene expression landscape and tumor representation. *Nature Communications*, *14*(1), 5417. https://doi.org/10.1038/s41467-023-41132-w

Jonsson, L. (2025). *Lejon/T-SNE-Java* [Java]. https://github.com/lejon/T-SNE-Java (Original work published 2014)

Katarey, D., & Verma, S. (2016). Drug-induced liver injury. *Clinical Medicine*, *16*(6, Supplement), s104–s109. https://doi.org/10.7861/clinmedicine.16-6-s104

Kim, S. (2021). Exploring Chemical Information in PubChem. *Current Protocols*, *1*(8), e217. https://doi.org/10.1002/cpz1.217

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2022). PubChem 2023 update. *Nucleic Acids Research*, *51*(D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956

KNIME | H2O.ai. (n.d.). Retrieved September 8, 2025, from https://h2o.ai/partner-network/find-a-partner/knime/

Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N. E. (Lucy), Strawbridge, S. A., Garcia-Patino, M., Kruger, R., Sivakumaran, A., Sanford, S., Doshi, R., Khetarpal, N., Fatokun, O., Doucet, D., Zubkowski, A., … Wishart, D. S. (2024). DrugBank 6.0: The DrugBank Knowledgebase for 2024.

Nucleic Acids Research, *52*(D1), D1265–D1275. https://doi.org/10.1093/nar/gkad976

Kolaric, T. O., Nincevic, V., Kuna, L., Duspara, K., Bojanic, K., Vukadin, S., Raguz-Lucic, N., Wu, G. Y., & Smolic, M. (2021). Drug-induced Fatty Liver Disease: Pathogenesis and Treatment. *Journal of Clinical and Translational Hepatology*, *9*(5), 731–737. https://doi.org/10.14218/JCTH.2020.00091

Korver, S., Bowen, J., Pearson, K., Gonzalez, R. J., French, N., Park, K., Jenkins, R., & Goldring, C. (2021). The application of cytokeratin-18 as a biomarker for drug-induced liver injury. *Archives of Toxicology*, *95*(11), 3435–3448. https://doi.org/10.1007/s00204-021-03121-0

Kubickova, B., & Jacobs, M. N. (2023). Development of a reference and proficiency chemical list for human steatosis endpoints in vitro. *Frontiers in Endocrinology*, *14*. https://doi.org/10.3389/fendo.2023.1126880

Li, X., Tang, J., & Mao, Y. (2022). Incidence and risk factors of drug-induced liver injury. *Liver International*, *42*(9), 1999–2014. https://doi.org/10.1111/liv.15262

Lichtenstein, D., Mentz, A., Schmidt, F. F., Luckert, C., Buhrke, T., Marx-Stoelting, P., Kalinowski, J., Albaum, S. P., Joos, T. O., Poetz, O., & Braeuning, A. (2020). Transcript and protein marker patterns for the identification of steatotic compounds in human HepaRG cells. *Food and Chemical Toxicology*, *145*, 111690. https://doi.org/10.1016/j.fct.2020.111690

Lin, J., Li, M., Mak, W., Shi, Y., Zhu, X., Tang, Z., He, Q., & Xiang, X. (2022). Applications of In Silico Models to Predict Drug-Induced Liver Injury. *Toxics*, *10*(12), 788. https://doi.org/10.3390/toxics10120788

List_fingerprints.pdf. (n.d.). Retrieved August 2, 2025, from https://www.pingzhang.net/bak/drugreposition/list_fingerprints.pdf

Liu, A., Walter, M., Wright, P., Bartosik, A., Dolciami, D., Elbasir, A., Yang, H., & Bender, A. (2021). Prediction and mechanistic analysis of drug-induced liver injury (DILI) based on chemical structure. *Biology Direct*, *16*, 6. https://doi.org/10.1186/s13062-020-00285-0

López-Pascual, E., Rienda, I., Perez-Rojas, J., Rapisarda, A., Garcia-Llorens, G., Jover, R., & Castell, J. V. (2024). Drug-Induced Fatty Liver Disease (DIFLD): A Comprehensive Analysis of Clinical, Biochemical, and Histopathological Data for Mechanisms Identification and Consistency with Current Adverse Outcome Pathways. *International Journal of Molecular Sciences*, *25*(10), Article 10. https://doi.org/10.3390/ijms25105203

Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., Petryszak, R., Ragueneau, E., Rothfels, K., Sevilla, C., Shamovsky, V., Stephan, R., Tiwari, K., Varusai, T., Weiser, J., … D'Eustachio, P. (2024). The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, *52*(D1), D672–D678. https://doi.org/10.1093/nar/gkad1025

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science*, *4*(5), 468–481. https://doi.org/10.1002/wcms.1183

Neuschwander-Tetri, B. A., Loomba, R., Sanyal, A. J., Lavine, J. E., Natta, M. L. V., Abdelmalek, M. F., Chalasani, N., Dasarathy, S., Diehl, A. M., Hameed, B., Kowdley, K. V., McCullough, A., Terrault, N., Clark, J. M., Tonascia, J., Brunt, E. M., Kleiner, D. E., & Doo, E. (2015). Farnesoid X nuclear receptor ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): A multicentre, randomised, placebo-controlled trial. *The Lancet*, *385*(9972), 956–965. https://doi.org/10.1016/S0140-6736(14)61933-4

Orlov, A., Akhmetshin, T., Horvath, D., Marcou, G., & Varnek, A. (2024). *From High Dimensions to Human Comprehension: Exploring Dimensionality Reduction for Chemical Space Visualization*. ChemRxiv. https://doi.org/10.26434/chemrxiv-2024-pxb4f

Ortega-Vallbona, R., Palomino-Schätzlein, M., Tolosa, L., Benfenati, E., Ecker, G. F., Gozalbes, R., & Serrano-Candelas, E. (2024). Computational Strategies for Assessing Adverse Outcome Pathways: Hepatic Steatosis as a Case Study. *International Journal of Molecular Sciences*, *25*(20), Article 20. https://doi.org/10.3390/ijms252011154

Paech, F., Bouitbir, J., & Krähenbühl, S. (2017). Hepatocellular Toxicity Associated with Tyrosine Kinase Inhibitors: Mitochondrial Damage and Inhibition of Glycolysis. *Frontiers in Pharmacology*, *8*. https://doi.org/10.3389/fphar.2017.00367

Retrieve/ID mapping | UniProt. (n.d.). Retrieved September 8, 2025, from https://www.uniprot.org/id-mapping

Schinderle, J. D., Wu, A., & Bochkis, I. M. (2024). *Reduced ZMPSTE24 expression leads to prelamin accumulation and development of steatosis in MASLD patients* (p. 2024.05.16.594546). bioRxiv. https://doi.org/10.1101/2024.05.16.594546

Sharma, B., & John, S. (2025). Nonalcoholic Steatohepatitis (NASH). In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK470243/

Shin, H. K., Kang, M.-G., Park, D., Park, T., & Yoon, S. (2020). Development of Prediction Models for Drug-Induced Cholestasis, Cirrhosis, Hepatitis, and Steatosis Based on Drug and Drug Metabolite Structures. *Frontiers in Pharmacology*, *11*. https://doi.org/10.3389/fphar.2020.00067

Sosnin, S. (2024). MolCompass: Multi-tool for the navigation in chemical space and visual validation of QSAR/QSPR models. *Journal of Cheminformatics*, *16*(1), 98. https://doi.org/10.1186/s13321-024-00888-z

Sosnin, S. (2024). *Sergsb/KNIME-standardizer* [Python]. https://github.com/sergsb/KNIME-standardizer (Original work published 2024)

Thakkar, S., Li, T., Liu, Z., Wu, L., Roberts, R., & Tong, W. (2020). Drug-induced liver injury severity and toxicity (DILIst): Binary classification of 1279 drugs by human hepatotoxicity. *Drug Discovery Today*, *25*(1), 201–208. https://doi.org/10.1016/j.drudis.2019.09.022

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Mekki Berrada, L., Sahraoui, M., Young, T., Bellemare Pépin, A., El Khantour, C., Landry, M., Pascarella, A., Hadid, V., Combrisson, E., O'Byrne, J., & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253. https://doi.org/10.1016/j.neuroimage.2023.120253

Tripp, A., Bacallado, S., Singh, S., & Hernández-Lobato, J. M. (2023, November 2). *Tanimoto Random Features for Scalable Molecular Machine Learning*. Thirty-seventh Conference on Neural Information Processing Systems. https://openreview.net/forum?id=MV0INFAKGq

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science*, *22*(1), 69–77. https://doi.org/10.1002/qsar.200390007

T-SNE (L. Jonsson). (n.d.). KNIME Community Hub. Retrieved September 8, 2025, from https://hub.knime.com/knime/extensions/org.knime.features.stats2/latest/org.knime.ext.tsne.node.TsneNodeFactory

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

Vrijenhoek, N. G., Wehr, M. M., Kunnen, S. J., Wijaya, L. S., Callegaro, G., Moné, M. J., Escher, S. E., & Water, B. van de. (2022). Application of high-throughput transcriptomics for mechanism-based biological read-across of short-chain carboxylic acid analogues of valproic acid. *ALTEX - Alternatives to Animal Experimentation*, *39*(2), Article 2. https://doi.org/10.14573/altex.2107261

Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F., Whaley, R., & Klein, T. E. (2021). An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clinical Pharmacology & Therapeutics*, *110*(3), 563–572. https://doi.org/10.1002/cpt.2350

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D. M., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, *52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004

Zhou, Y., Zhang, Y., Zhao, D., Yu, X., Shen, X., Zhou, Y., Wang, S., Qiu, Y., Chen, Y., & Zhu, F. (2024). TTD: Therapeutic Target Database describing target druggability

information. *Nucleic Acids Research*, *52*(D1), D1465–D1477. https://doi.org/10.1093/nar/gkad751

Zhu, L., Yang, X., Wu, S., Dong, R., Yan, Y., Lin, N., Zhang, B., & Tan, B. (2024). Hepatotoxicity of epidermal growth factor receptor—Tyrosine kinase inhibitors (EGFR-TKIs). *Drug Metabolism Reviews*, *56*(3), 302–317. https://doi.org/10.1080/03602532.2024.2388203

Zhu, X., & Kruhlak, N. L. (2014). Construction and analysis of a human hepatotoxicity database suitable for QSAR modeling using post-market safety data. *Toxicology*, *321*, 62–72. https://doi.org/10.1016/j.tox.2014.03.009