Toxicology Letters 381 (2023) 20–26

Contents lists available at ScienceDirect

Toxicology Letters

journal homepage: www.journals.elsevier.com/toxicology-letters

Toxicity prediction using target, interactome, and pathway profiles as descriptors

Barbara Füzi [a], Neann Mathai [b], Johannes Kirchmair [a,b], Gerhard F. Ecker [a,*]

[a] Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria
[b] Department of Chemistry and Computational Biology Init (CBU), University of Bergen, N-5020 Bergen, Norway

ARTICLE INFO

Editor: Dr. Angela Mally

Keywords:
Prediction
Machine learning
Pathway
Target
Hepatotoxicity

ABSTRACT

In silico methods are essential to the safety evaluation of chemicals. Computational risk assessment offers several approaches, with data science and knowledge-based methods becoming an increasingly important sub-group. One of the substantial attributes of data science is that it allows using existing data to find correlations, build strong hypotheses, and create new, valuable knowledge that may help to reduce the number of resource intensive experiments.

In choosing a suitable method for toxicity prediction, the available data and desired toxicity endpoint are two essential factors to consider. The complexity of the endpoint can impact the success rate of the *in silico* models. For highly complex endpoints such as hepatotoxicity, it can be beneficial to decipher the toxic event from a more systemic point of view.

We propose a data science-based modelling pipeline that uses compounds` connections to tissue-specific biological targets, interactome, and biological pathways as descriptors of compounds. Models trained on different combinations of the collected, compound-target, compound-interactor, and compound-pathway profiles, were used to predict the hepatotoxicity of drug-like compounds. Several tree-based models were trained, utilizing separate and combined target, interactome and pathway level variables. The model using combined descriptors of all levels and the random forest algorithm was further optimized. Descriptor importance for model performance was addressed and examined for a biological explanation to define which targets or pathways can have a crucial role in toxicity. Descriptors connected to cytochromes P450 enzymes, heme degradation and biological oxidation received high weights. Furthermore, the involvement of other, less discussed processes in connection with toxicity, such as the involvement of RHO GTPase effectors in hepatotoxicity, were marked as fundamental. The optimized combined model using only the selected descriptors yielded the best performance with an accuracy of 0.766.

The same dataset using classical Morgan fingerprints for compound representation yielded models with similar performance measures, as well as the combination of systems biology-based descriptors and Morgan fingerprints. Consequently, adding the structural information of compounds did not enhance the predictive value of the models. The developed systems biology-based pipeline comprises a valuable tool in predicting toxicity, while providing novel insights about the possible mechanisms of the unwanted events.

1. Introduction

In silico methods have matured into key methods for the safety assessment of drugs. They can detect early on the possibility of an undesired event and, compared to animal testing, are less costly, more time-efficient, and free of ethical concerns Raies, Bajic (2016). Given their benefits, many methods are being developed in this area, supporting different tasks of the safety assessment pipeline.

Two critical factors to consider in choosing the most accurate method are the availability of the data and the endpoint of interest. For instance, for integrating and modelling mechanistic data, physiologically based pharmacokinetic (PBPK) modelling can be beneficial

Abbreviations: RF, Random Forest; GBT, Gradient Boosted Tree; DT, Decision Tree; WF, workflow; ML, machine learning; FDA, Food and Drug Administration; ID, identifier; SPM, specialized proresolver mediator; INDVAL, indicator values.

* Corresponding author.
E-mail address: gerhard.f.ecker@univie.ac.at (G.F. Ecker).

https://doi.org/10.1016/j.toxlet.2023.04.005
Received 26 September 2022; Received in revised form 11 March 2023; Accepted 12 April 2023
Available online 13 April 2023

(Lipscomb et al., 2012). Meanwhile, machine learning-based regression and classification models typically make use of the structural information of chemical compounds. These models usually learn from the data via an algorithm to help classify compounds into binary groups. A desirable dataset for this purpose contains many data points of each class. The compounds` representation for the learning algorithms can be manifold. Typically, the possible descriptors of the compounds are molecular descriptors and/or physicochemical properties.

In addition to data availability, the complexity of the endpoint can have a crucial impact on the success rate of the machine learning model. Predicting the inhibition of a single off-target, such as hERG, typically is a more straightforward task than predicting a complex endpoint such as hepatotoxicity (Minerali et al., 2020). Established machine learning methods based on algorithms such as random forest or support vector machine are frequently employed to predict toxicity of drugs with tree-based models being particularly useful for classification tasks (Mienye et al., 2019). Furthermore, also in the area of toxicity prediction, deep learning approaches are becoming increasingly popular (Liu et al., 2022).

In addressing complex endpoints, it is worth considering utilizing the broader effect of drugs on the human organism or tissue of interest. Systems biology can address the holistic aspects of interactions between compounds and organisms, by deciphering biological entities and mechanisms being involved. It integrates experimental data for creating novel insights. Systems biology is being used in toxicology, for example in the field of toxicogenomics, developing adverse outcome pathways or supporting network-based toxicology approaches (Bugrim et al., 2004).

Here, a new approach is introduced, where the impact of compounds on target, interactome and pathways levels are being used for the representation of drugs. This predictive modelling method employs drugs` systemic fingerprints as descriptors for the training of machine learning models. Three layers of interacting biological entities (target, interactome, pathway) were included to estimate the effect of drugs on different biological complexity levels by using these events as unique descriptors for each compound. The machine learning pipeline was built in KNIME, utilizing native, tree based KNIME machine learning nodes.

2. Methods

2.1. Software tools

KNIME analytics is a platform for creating data science workflows, whereas H2O.ai offers a wide range of AI solutions. Both platforms together created the KNIME H2O Driverless AI Extension, an integration of H2O Driverless AI in KNIME (Combining the power of KNIME and H2O.ai in a single integrated workflow [wen-document]). In addition to native KNIME nodes, nodes from H2O.ai, were used in our modelling pipeline. The H2O nodes enable the creation of flexible, mix-and-match automated machine learning pipelines, which are easy to understand and of high quality.

2.2. Source of DILI data

The base dataset for the model building was the Drug Induces Liver Injury (DILI) Rank dataset provided by the FDA (Chen et al., 2016). The dataset was processed as described in detail in a previous publication (Füzi et al., 2022). DILIRank is based on drug labelling (Chen et al., 2011), the LiverTox database of the National Institute of Health (Hoofnagle et al., 2013), Drug-Induced Liver Injury-Network Studies (Fontana et al., 2009), and publications such as from Suzuki et al. (Suzuki et al., 2010). There are 4 categories available in this dataset according to their abilities causing DILI – mostDILIL, lessDILI, ambiguousDILI and noDILI. The grouping process takes into consideration if there is a clear verification available of the drug causing DILI. Our analysis was carried out with the mostDILI and noDILI groups.

Ultimately, the toxic group consists of 180 compounds with high DILI concern and of 272 compounds without any DILI concern.

2.3. Descriptors

For validating if only based on information of the systemic effect of the drugs we can predict toxicity, structural properties of the compounds were not considered as descriptors for the ML models in the first evaluation of the method. Instead, a previously developed, openly available KNIME workflow (https://kni.me/w/Yf0V_0m1wm0Sw7O) (Füzi et al., 2021) was utilized for creating target and pathway fingerprints for the compounds. In this context, targets are all proteins on which a compound is active, and pathways are biological processes that result in a product (such as new molecules) or change (for example metabolic transformation) in the cell (Biological Pathways Fact Sheet [web-document]).

Via the KNIME workflow, data in openly available repositories was retrieved, curated, and integrated. The repositories are summarized in Table 1.

The output is a list of target proteins and biological pathways that can be connected to the compounds. Furthermore, the collected target list was filtered for liver-specific proteins using the Human Protein Atlas database (Thul and Lindskog, 2018) since the analysed compounds are part of a hepatotoxicity dataset and the aim was to predict hepatotoxicity. The detailed process of creating target and pathway profiles of compounds is summarized visually in Fig. 1 and described in detail previously (Füzi et al., 2021).

The created list of the compounds` connections to targets, tissue-specific targets and pathways were translated into binary matrices representing the connectivity between compound and biological entity. 1 stands for an existing connection, while 0 means not existing/not known connections.

For the machine learning project, the approach summarized in Fig. 1 was complemented with predicted targets and interactome data.

In addition to the protein targets retrieved from openly available, high quality databases (Fig. 1), additional predicted targets for the mostDILI and noDILI compounds were also retrieved using a validated similarity-based target prediction approach with a large target coverage Mathai, Kirchmair (2020) Reference data for the target prediction was curated from the ChEMBL database (version 27) to retrieve high quality bioactivity data. The ChEMBL compounds were standardized and unique compound-protein pairs with an activity value less than or equal to 10,000 nM were retained as the reference set for the target prediction (Mathai et al., 2021). Compounds were encoded using ECFP2 fingerprints for the prediction model. The DILI compounds were standardized using the same pipeline as the reference set, the Tanimoto similarity between the compounds of interest and the reference compounds was

Table 1
Utilized repositories and their accessibility.

Database	Type of data used	Accessibility	Reference
ChEMBL	compound-target	programmatic access	(Davies et al., 2015)
DrugBank	compound-target	download	(Wishart et al. (2018))
IUPHAR	compound-target	programmatic access	(Armstrong et al., 2020)
TTD	compound-target	download	(Wang et al., 2020)
PharmGKB	compound-target	download	Whirl-Carrillo et al. (2012)
UniProt	protein identifiers	programmatic access	(Consortium, 2019)
Human Protein Atlas	tissue expression	download	(Uhlén et al., 2015)
Reactome	pathway	programmatic access	(Jassal et al., 2020)

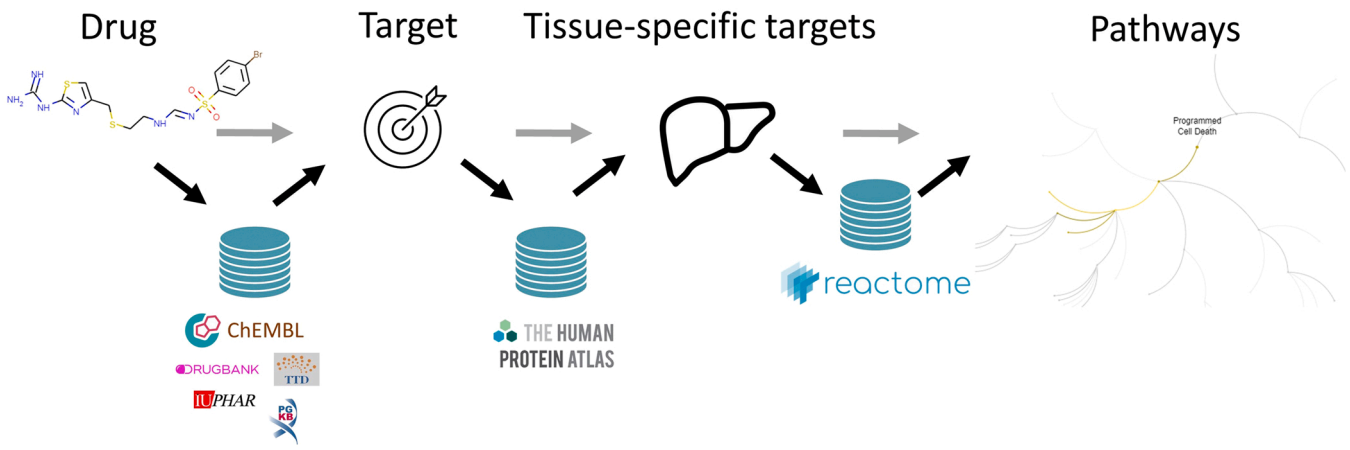

Fig. 1. Graphical representation of building the target and pathway profiles of the compounds.

calculated, and the compounds of interest were assigned the reference compounds' targets when the similarity was 0.5 or greater. The predicted targets were added to the retrieved known target list, via translating the CHEMBL IDs to UniProt IDs and creating compound target pairs of each set of data. If the predicted interaction conflicted with known interactions, the latter was used.

Furthermore, an additional interactome layer was established. The interactome workflow searches first-degree interactor proteins of the direct target proteins in the MINT (Chatr-aryamontri et al., 2007) and IntAct (Hermjakob et al., 2004) databases. The interactors are human, single-type proteins with an interaction score of 0.5 or above. The interaction scores are obtained from the respective databases - MINT or INtAct - and indicate the confidence in the protein-protein interaction. This workflow is now openly accessible (https://hub.knime.com/barbaraf/spaces/Public/latest/~7jbMNHvalhE2ZCtU/) and a use case for proteomics analysis is also available (Rodrigues et al., 2022).

The output of the fingerprint creation consists of 4 binary matrices representing the connection between DILI compounds and different levels of biological entities and organizations: compound-target, compound-target+predicted target, compound-interactome and compound-pathway.

2.4. Machine learning models

To evaluate the possible performance range of the approach and select the best algorithm for the task, 4 * 3 sub-workflows were created using four different types of fingerprints and three different model algorithms. The fingerprints used were: compound-target, compound-target+predicted target, compound-interactome, and compound-pathway fingerprints, separately. For creating the models, the corresponding compound connection table (supplementary files: DILI_workflow_targets, DILI_predicted_targets, DILI_interactome, DILI_pathways) was read into the workflow. Three different tree-based machine learning algorithms used were: random forest (RF) (Breiman, 2001), gradient boosted tree model (GBT) Friedman (2001), decision tree model (DT) (Myles et al., 2004).

Since the workflow is repeated for all four datasets, here, only an overview is given based on the compound-target dataset. A matrix of compounds (rows) vs. targets (columns) was generated with either a 1, signifying an interaction between the compound and target, or a 0, indicating a non-interacting or unknown compound-target pair. Furthermore, the compound IDs were matched with the toxic and non-toxic compound lists, creating a new string class column, indicating if the compound is "toxic" or "non-toxic". Model building was performed with the created table as the input, including 5-fold cross-validation using X-partitioner and Y-aggregator nodes. Three different classification models, one per algorithm, were trained and tested to classify compounds as toxic or non-toxic. With the Scorer node, model performance was evaluated by calculating the mean specificity, sensitivity, precision, accuracy, and the standard deviation (SD) of the accuracy of the models. Fig. 2.

2.5. Combination

Based on the results of the separate models, a combined model utilizing the target, interactome, and pathway profiles was also established. The integrated table included those compounds which had data in all three profiles. This table has 2995 columns – 2993 descriptors-, and 295 rows -compounds- (supplementary file: DILI_combined). The rest of the pipeline was carried out as described above.

2.6. Descriptor importance

The first models were built with 2993 descriptors, of which 387 were target, 1572 interactome and 1034 pathway-based. Use of highly correlated or irrelevant descriptors can weaken the robustness of a model (Skoraczyński et al., 2017). Therefore, an evaluation was performed on the combined model to find the most significant descriptors. The data was fed into the H2O random forest node to utilize the variable importance measure possibility of the H2O framework. The feature importance was determined by calculating the relative and scaled influence of each variable. The deciding factors of the calculation are whether the variable was selected as a decision node of the decision tree and if yes how much the squared error decreased. The exact calculation process can be found in the H2O documentation ("Variable Importance — H2O 3.36.1.2 documentation").

All variables, proteins, and pathways were analysed simultaneously. The cut-off for the relative importance was set to 30 or above, correlating to a scaled influence of at least 0.1, which indicates that the feature is to at least some extent important. Consequently, the models were retrained with 72 descriptors.

2.7. Optimization

Using the Parameter Optimization loop, variables for tree depth and number of trees were created, 10–40; 50–200 as the start -and stop values, 5 and 50 as step size, respectively. The search strategy was set to brute force (Ababneh et al., 2006); therefore, all possible parameter combinations were checked. The Parameter Optimization loop was combined with cross-validation to prevent overtraining.

2.8. Morgan fingerprints

To address the predictive value of the developed method, baseline

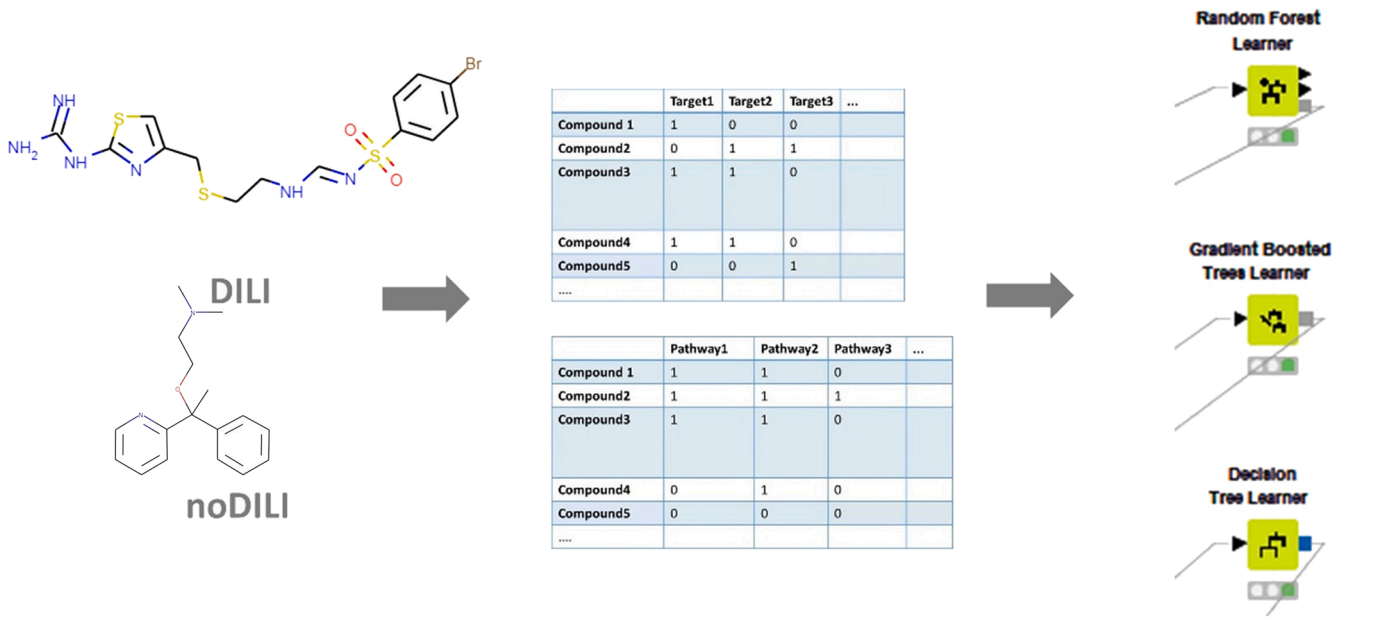

Fig. 2. Schematic representation of the Machine Learning pipeline.

machine learning models using Morgan fingerprints, as well as models utilizing a fusion of both Morgan fingerprints and the introduced systemic fingerprints were trained analogue to the models described above. To create Morgan fingerprints, first the UniChem API was queried to retrieve the corresponding InChI-keys for the ChEMBL IDs of the compounds (Chambers et al., 2013). After that the InChI keys were transferred to RDKit molecules and in the next step Morgan fingerprints using RDKit fingerprint node were created.

3. Results

3.1. Compound dataset

The initial DILI compound dataset would be to some extent imbalanced. However, target, interactome or pathway data could not be found for every compound. Table 1 presents the distribution of the actual input data after building the corresponding target, target + predicted target, interactome, pathway profiles, and the combination table. Table 2.

Table 2 indicates that the base-datasets for the models are not 100% identical, due to limitations in data availability. However, the core of the datasets consists of the same compounds.

3.2. Model performances

Using different sets of the systemic descriptors, tree-based models were trained.

This part of the Result section summarizes the statistical measurements of the models created via the separated fingerprints. 387 target, 357 predicted target, 1572 interactome and 1034 pathway type descriptors were utilized for building the initial models without feature optimization. Table 3.

The GBT algorithm slightly outperformed the other two models in three of four cases. The DT models had the lowest overall scores.

Based on the results of the separated fingerprint models, one combined model using the target, interactome and pathway profiles as a combined fingerprint was created. The predicted targets were not used for this part of the exercise since including them slightly worsened the performance compared to the target fingerprint-based model. Table 4.

3.2.1. Combined fingerprint (target+interactome+pathway)

This first evaluation showed that the models performed similarly, with sensitivity being the lowest performance attribute.

3.3. Descriptor importance

After evaluating the feature importance, 72 descriptors from the pool of all descriptors with a relative importance 30 or above, correlating to a scaled influence of at least 0.1 were selected via the H2O variable importance measures to retrain the model. The top-10 descriptors are reported in Table 5 whereas the full list is provided as supplement (supplementary file: top_descriptors).

There are 21 target, 5 interactome, and 46 pathway related descriptors among the most significant ones. Furthermore, the Indicator Value (INDVAL) was calculated, to provide an additional, informative relevance score (Dufrêne and Legendre, 1997). INDVAL implies, how relevant is a descriptor for a group (here group 1: non-toxic, group 2: toxic). For calculating the INDVAL, the R package "labdsv" was used. This scoring found once again biological oxidation, metabolism- and in addition specialized proresolving meadiators (SPMs) - related descriptors the most relevant for the toxic class. The results of this are summarized in the supplementary file indval_results.csv, with a 0.05 cut-off for the p-value to include statistically relevant findings, including frequencies showing how often the variable was present among samples (here rows).

3.4. Retraining the model

The performance statistics of the retrained model using only the 72 selected features as descriptors and the two previously best-performing algorithms (RF and GBT) is summarized in Table 6.

Table 2
Number of compounds in the base datasets for the machine learning models.

Descriptors	Positive (DILI yes)	Negative (DILI no)
Target	150	176
Target+Predicted target	163	195
Interactome	143	155
Pathway	148	175
Combined	141	154
Morgan Fingerprint	170	200
Morgan Fingerprint + Combined	135	147

Table 3
Model performance of different descriptor sets in combination with different algorithms – means of results of 5-fold cross validation with 80/20 split.

Descriptor Set	Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	Accuracy (SD)
Target	RF	0.659	0.535	0.779	0.675	0.053
Target	**GBT**	**0.718**	**0.570**	**0.799**	**0.690**	**0.051**
Target	DT	0.645	0.569	0.730	0.653	0.042
Target+Predicted Target	**RF**	**0.652**	**0.530**	**0.752**	**0.651**	**0.066**
Target+Predicted Target	GBT	0.639	0.508	0.763	0.642	0.046
Target+Predicted Target	DT	0.606	0.556	0.701	0.640	0.069
Interactome	RF	0.67	0.418	0.809	0.617	0.069
Interactome	**GBT**	**0.707**	**0.603**	**0.757**	**0.678**	**0.066**
Interactome	DT	0.621	0.562	0.684	0.621	0.028
Pathway	RF	0.618	0.628	0.668	0.65	0.028
Pathway	**GBT**	**0.688**	**0.615**	**0.759**	**0.694**	**0.039**
Pathway	DT	0.593	0.570	0.678	0.625	0.021

Table 4
Combined model performances – means of results of 5-fold cross validation with 80/20 split.

Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	Accuracy (SD)
RF	**0.668**	**0.632**	**0.71**	**0.675**	**0.066**
GBT	0.665	0.588	0.725	0.658	0.06
DT	0.605	0.545	0.677	0.614	0.05

Table 5
Top 10 descriptors selected in the feature importance exercise.

Descriptor	**Type**
Fatty acid metabolism	pathway
Arachidonic acid metabolism	pathway
Heme degradation	pathway
Biological oxidations	pathway
Metabolism of porphyrins	pathway
RHO GTPase Effectors	pathway
P11712 - Cytochrome P450 2C9	target
O95870 - Phosphatidylserine lipase ABHD16A	target
Immune System	pathway
Phase II - Conjugation of compounds	pathway

3.5. Optimization

The best-performing model (72 combined descriptors, RF) was further optimized. Table 6 represents the results of parameter optimization combined with cross-validation, including the best parameter combinations. The variance across the folds was 0.055. Table 7.

3.6. Morgan Fingerprint

For comparing the predictive value of the developed method with traditional models based on standard chemical descriptors, additional models using classical fingerprints were trained. Based-on InChI-Keys 170 DILI and 210 no DILI compounds were represented via Morgan fingerprints in a slightly imbalanced dataset for model training. The results of the model performance are presented in Table 8.

The last set of models were trained with both Morgan fingerprints and the above introduced systemic descriptors, introducing a unique combination of representing molecules in a machine learning process. For 147 non-toxic and 135 toxic compounds also, systemic descriptors were available. Table 9.

The RF model was further optimized, with variables tree depth and number of trees tested in combination via brute force.Table 10.

4. Discussion

Investigating and predicting liver toxicity is a complex task. In this work, a novel, systems biology-based approach was introduced, where the compound dataset was presented by target, interactome, and pathway profiles as descriptors.

Using the H20 framework for descriptor selection 72 descriptors were selected based on their importance score for retraining the model (supplementary file: top_descriptors). To create hypotheses of likely mechanisms of toxicity, the top ranked descriptors (Table 4) were compared to literature for existing evidence and explanations of their possible role in hepatotoxicity: Fatty acid metabolism was linked to valproate acid-induced hepatotoxicity (Ji et al. (2010), arachidone acid metabolism can be connected to inflammation (Higgins and Lees, 1984), and of course, to anti-inflammatory drugs. The induction of heme oxygenase-1 expression has been shown to protect the liver from injuries induced by several xenobiotics (Origassa and Câmara, 2013). Oxidative stress-mediated hepatotoxicity is a known process where free radicals damage the components of the cell (Videla et al., 2003); therefore, the high ranking as a descriptor of the pathway Biological Oxidation is not unexpected. The Rho GTPase effectors are AGC-family serine/threonine kinases involved in a diverse range of cellular processes of cell survival (Clayton and Ridley, 2020). Finally, CYP Enzymes are well-known for

Table 7
Performance after parameter optimization.

Precision	Sensitivity	Specificity	Accuracy	treedepth	numtrees
0.769	**0.730**	**0.799**	**0.766**	**30**	**190**
0.765	**0.738**	**0.792**	**0.766**	**48**	**460**

Table 8
Model performance with Morgan Fingerprints after 5-fold cross-validation.

Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	Accuracy (SD)
RF	0.58	0.641	0.638	0.636	0.049
GBT	0.672	0.641	0.749	0.703	0.039

Table 6
Performance statistics,retrained models – means of results of 5-fold cross validation with 80/20 split.

Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	Accuracy (SD)	Cohen`s Kppa
RF	**0.729**	**0.716**	**0.762**	**0.736**	**0.07**	**0.471**
GBT	0.71	0.607	0.776	0.695	0.062	0.377

Table 9
Model performance using both Morgan fingerprints and the systemic descriptors (72) - means of results of 5-fold cross validation with 80/20 split.

Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	Accuracy (SD)
+RF	0.713	0.719	0.742	0.723	0.05
+GBT	0.744	0.677	0.776	0.713	0.034

Table 10
Performance of the best model using both Morgan fingerprints and systemic descriptors as the representation of compounds - means of results of 5-fold cross validation with 80/20 split.

Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	Accuracy (SD)
+oRF	0.768	0.715	0.8	0.759	0.048

their involvement in drug metabolism (Villeneuve and Pichette, 2004). In particular, Cytochrome P450 2C9 has been associated with hepatotoxicity caused by valproic acid (Ho et al., 2003) (Zhao et al., 2017). Comparing our results with metabolism studies, possible mode of toxicities could be found, where the found descriptors can have a crucial role. For example changes in phospholipid metabolism Liu et al. (2022), disruption in long chain fatty acid mechanism (Ramirez et al., 2018). Cytochrome P450-metiated metabolism was categorized as obligatory for inducing hepatic toxicity by some substances (Pandit, 2012). It is inevitable that metabolism plays a crucial role in a huge portion of liver toxic events caused by drugs.

Additionally, calculating the indicator value yielded the information that specialized proresolving mediators (SPMs) appear to be also relevant in toxic events. These mediators have key roles in the metabolic homeostasis of the liver, and in avoiding inflammation (Musso et al., 2018). All the findings serve as a base for understanding the involvement of the biological entities in the event of hepatotoxicity and provide some information about the type of drugs causing liver injuries.

The most important descriptors came mainly from the pathway category, less from the target category, and least from the interactome category. This underlines the importance of drugs` connections to not only direct targets but to biological pathways in estimating the effects of the compounds.

Without the descriptor selection, the models were performing significantly lower, especially regarding their sensitivity. In general, this approach relies strongly on publicly available data, which is generally sparse. With adding predicted targets, the performance has dropped. This indicates that the signal for toxicity was lost in the huge number of descriptors, missing data and added noise. Therefore, it can be estimated that a denser matrix of experimentally proven values could yield better performance, and the predictive models would benefit from a set of complete or closely complete target, interactome, and pathway profiles.

After utilizing novel descriptors, models with classical Morgan fingerprints were trained. The GBT algorithm performed slightly better and the RF model slightly worse with these fingerprints, than the separated systemic models. Combining Morgan fingerprints with the 72 most important systemic descriptors yielded a better performing model than the one trained only with Morgen fingerprints, however still the systemic model with the selected descriptors performed best. Optimizing both models by parameter optimization, similar accuracy could be achieved. With our descriptor selection pipeline, we aimed to have biological insights of the process of liver toxicity, as well as reducing the "black-box" characteristics of our machine learning models. Optimizing a machine learning model is an important step for achieving accurate predictions but also for using the computational resources wisely, however creating biological insight with these optimization steps is not straight-forward.

Interestingly, the model's performance was not enhanced by adding the structural information of the molecules to the matrix. However, combining the systemic information with other methods, such as expression networks or deep learning can be beneficial. However, it is important to mention that retrieving target and pathway information based on activity data is more complex, than using the compound`s molecular information. In toxicity prediction for novel compounds, where there is no openly available activity data in public repositories is available, it is advisable to use our pipeline in combination with similarity-based approaches. In that case, target, interactome and pathway profiles of similar compounds to the compound of interest can be retrieved and merged, based on Tanimoto similarity for instance, to create a unique profile for novel compounds and use it further in toxicity prediction.

5. Conclusion

This study demonstrates a different way of training predictive models by using systemic information, with robust performance measures. Additionally, investigating the most important biological entities and processes for the predictive models could yield a novel insight into possible mechanisms of hepatotoxic events.

Funding

This work has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement number 116030 (TransQST). This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Programme and European Federation of Pharmaceutical Industries and Associations (EFPIA). This work was also supported by the Austrian Science Fund/FWF, grant W1232 (MolTag).

CRediT authorship contribution statement

GE and BF contributed to the conceptualization of the study. BF created the machine learning pipeline and performed the analysis of the results. JK and NM developed and applied the machine learning models for target prediction. BF wrote the first version of the manuscript, GE, JK, and NM contributed to the manuscript refinement and editing. Authors declare that this work reflects only the author's view, and IMI-JU is not responsible for any use that may be made of the information it contains. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data is available in the supplementary material.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.toxlet.2023.04.005.

References

Ababneh, M., Oqeili, S., Abdeen, R., 2006. Occurrences algorithm for string searching based on brute-force algorithm. J. Comput. Sci. 2 https://doi.org/10.3844/jcssp.2006.82.85.

Armstrong, J.F., Faccenda, E., Harding, S.D., Pawson, A.J., Southan, C., Sharman, J.L., Campo, B., Cavanagh, D.R., Alexander, S.P.H., Davenport, A.P., Spedding, M., Davies, J.A., NC-IUPHAR, 2020. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV

Guide to MALARIA PHARMACOLOGY. Nucleic Acids Res 48, D1006–D1021. https://doi.org/10.1093/nar/gkz951.
Biological Pathways Fact Sheet [web-document], Genome.gov. URL: ⟨https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet⟩ (Accessed 7 March 22).
Breiman, L., 2001. Random Forests. Mach. Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Bugrim, A., Nikolskaya, T., Nikolsky, Y., 2004. Early prediction of drug metabolism and toxicity: systems biology approach and modeling. Drug Discov. Today 9, 127–135. https://doi.org/10.1016/S1359-6446(03)02971-4.
Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., Overington, J.P., 2013. UniChem: a unified chemical structure cross-referencing and identifier tracking system. J. Chemin.-. 5, 3. https://doi.org/10.1186/1758-2946-5-3.
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 Res 35, D572–D574. https://doi.org/10.1093/nar/gkl950.
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 Discov. Today 21, 648–653. https://doi.org/10.1016/j.drudis.2016.02.015.
Chen, M., Vijay, V., Shi, Q., Liu, Z., Fang, H., Tong, W., 2011. FDA-approved drug labeling for the study of drug-induced liver injury. Drug Discov. Today 16, 697–703. https://doi.org/10.1016/j.drudis.2011.05.007.
Clayton, N.S., Ridley, A.J., 2020. Targeting Rho GTPase signaling networks in cancer. Front. Cell Dev. Biol. 8.
Combining the power of KNIME and H2O.ai in a single integrated workflow [wen-document], KNIME. URL: ⟨https://www.knime.com/blog/combining-the-power-of-knime-and-h2oai-in-a-single-integrated-workflow⟩ (accessed 2.4.22).
Consortium, T.U., 2019. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47, D506–D515. https://doi.org/10.1093/nar/gky1049.
Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., Overington, J.P., 2015. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 43, W612–W620. https://doi.org/10.1093/nar/gkv352.
Dufrêne, M., Legendre, P., 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecol. Monogr. 67, 345–366. https://doi.org/10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2.
Fontana, R.J., Watkins, P.B., Bonkovsky, H.L., Chalasani, N., Davern, T., Serrano, J., Rochon, J., DILIN Study Group, 2009. Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. Drug Saf. 32, 55–68.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.
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. Front. Pharm. 12.
Füzi, B., Malik-Sheriff, R.S., Manners, E.J., Hermjakob, H., Ecker, G.F., 2022. KNIME workflow for retrieving causal drug and protein interactions, building networks, and performing topological enrichment analysis demonstrated by a DILI case study. J. Chemin.-. 14, 37. https://doi.org/10.1186/s13321-022-00615-6.
Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R., 2004. IntAct: an open source molecular interaction database. Nucleic Acids Res 32, D452–D455. https://doi.org/10.1093/nar/gkh052.
Higgins, A.J., Lees, P., 1984. The acute inflammatory process, arachidonic acid metabolism and the mode of action of anti-inflammatory drugs. Equine Vet. J. 16, 163–175. https://doi.org/10.1111/j.2042-3306.1984.tb01893.
Ho, P.C., Abbott, F.S., Zanger, U.M., Chang, T.K.H., 2003. Influence of CYP2C9 genotypes on the formation of a hepatotoxic metabolite of valproic acid in human liver microsomes. Pharm. J. 3, 335–342. https://doi.org/10.1038/sj.tpj.6500210.
Hoofnagle, J.H., Serrano, J., Knoben, J.E., Navarro, V.J., 2013. LiverTox: a website on drug-induced liver injury. Hepatology 57, 873–874. https://doi.org/10.1002/hep.26175.
Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., Loney, F., May, B., Milacic, M., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2020. The reactome pathway knowledgebase. Nucleic Acids Res. 48, D498–D503. https://doi.org/10.1093/nar/gkz1031.
Ji, Q., Shi, X., Lin, R., Mao, Y., Zhai, X., Lin, Q., Zhang, J., 2010. Participation of lipid transport and fatty acid metabolism in valproate sodium-induced hepatotoxicity in HepG2 cells. Toxicol. Vitr. 24, 1086–1091. https://doi.org/10.1016/j.tiv.2010.03.014.
Lipscomb, J.C., Haddad, S., Poet, T., Krishnan, K., 2012. Physiologically-based pharmacokinetic (PBPK) models in toxicity testing and risk assessment. Adv. Exp. Med. Biol. 745, 76–95. https://doi.org/10.1007/978-1-4614-3055-1_6.
Liu, J., Guo, W., Sakkiah, S., Ji, Z., Yavas, G., Zou, W., Chen, M., Tong, W., Patterson, T. A., Hong, H., 2022. Machine learning models for predicting liver toxicity. In: Benfenati, E. (Ed.), In Silico Methods for Predicting Drug Toxicity, Methods in Molecular Biology. Springer, US, New York, NY, pp. 393–415. https://doi.org/10.1007/978-1-0716-1960-5_15.
Liu, L., Wu, Q., Miao, X., Fan, T., Meng, Z., Chen, X., Zhu, W., 2022. Study on toxicity effects of environmental pollutants based on metabolomics: A review. Chemosphere 286, 131815. https://doi.org/10.1016/j.chemosphere.2021.131815.
Mathai, N., Kirchmair, J., 2020. Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope. Int. J. Mol. Sci. 21, 3585. https://doi.org/10.3390/ijms21103585.
Mathai, N., Stork, C., Kirchmair, J., 2021. BonMOLière: small-sized libraries of readily purchasable compounds, optimized to produce genuine hits in biological screens across the protein space. Int. J. Mol. Sci. 22, 7773. https://doi.org/10.3390/ijms22157773.
Mienye, I.D., Sun, Y., Wang, Z., 2019. Prediction performance of improved decision tree-based algorithms: a review. Procedia Manuf. 2nd Int. Conf. Sustain. Mater. Process. Manuf., SMPM 2019, 8-10 March 2019, Sun City, South Afr. 35, 698–703. https://doi.org/10.1016/j.promfg.2019.06.011.
Minerali, E., Foil, D.H., Zorn, K.M., Lane, T.R., Ekins, S., 2020. Comparing machine learning algorithms for predicting drug-induced liver injury (DILI. Mol. Pharm. 17, 2628–2637. https://doi.org/10.1021/acs.molpharmaceut.0c00326.
Musso, G., Gambino, R., Cassader, M., Paschetta, E., Sircana, A., 2018. Specialized proresolving mediators: enhancing nonalcoholic steatohepatitis and fibrosis resolution. Trends Pharmacol. Sci. 39, 387–401. https://doi.org/10.1016/j.tips.2018.01.003.
Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D., 2004. An introduction to decision tree modeling. J. Chemom.: A J. Chemom. Soc. 18 (6), 275–285.
Origassa, C.S.T., Câmara, N.O.S., 2013. Cytoprotective role of heme oxygenase-1 and heme degradation derived end products in liver injury. World J. Hepatol. 5 (10), 541.
Pandit, 2012. Drug-induced hepatotoxicity: a review. J. Appl. Pharm. Sci. https://doi.org/10.7324/JAPS.2012.2541.
Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. WIREs Comput. Mol. Sci. 6, 147–172. https://doi.org/10.1002/wcms.1240.
Ramirez, T., Strigun, A., Verlohner, A., Huener, H.-A., Peter, E., Herold, M., Bordag, N., Mellert, W., Walk, T., Spitzer, M., Jiang, X., Sperber, S., Hofmann, T., Hartung, T., Kamp, H., van Ravenzwaay, B., 2018. Prediction of liver toxicity and mode of action using metabolomics in vitro in HepG2 cells. Arch. Toxicol. 92, 893–906. https://doi.org/10.1007/s00204-017-2079-6.
Rodrigues, D., Coyle, L., Füzi, B., Ferreira, S., Jo, H., Herpers, B., Chung, S.-W., Fisher, C., Kleinjans, J.C.S., Jennen, D., de Kok, T.M., 2022. Unravelling mechanisms of doxorubicin-induced toxicity in 3D human intestinal organoids. Int. J. Mol. Sci. 23, 1286. https://doi.org/10.3390/ijms23031286.
Skoraczyński, G., Dittwald, P., Miasojedow, B., Szymkuć, S., Gajewska, E.P., Grzybowski, B.A., Gambin, A., 2017. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient. Sci. Rep. 7, 3582. https://doi.org/10.1038/s41598-017-02303-0.
Suzuki, A., Andrade, R.J., Bjornsson, E., Lucena, M.I., Lee, W.M., Yuen, N.A., Hunt, C.M., Freston, J.W., 2010. Drugs associated with hepatotoxicity and their reporting frequency of liver adverse events in VigiBase™. Drug Saf. 33, 503–522. https://doi.org/10.2165/11535340-000000000-00000.
Thul, P.J., Lindskog, C., 2018. The human protein atlas: a spatial map of the human proteome. Protein Sci. Publ. Protein Soc. 27, 233–244. https://doi.org/10.1002/pro.3307.
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., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J.M., Hamsten, M., Feilitzen, K., von, Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Heijne, G., von, Nielsen, J., Pontén, F., 2015. Tissue-based map of the human proteome. Science 347. https://doi.org/10.1126/science.1260419.
Variable Importance — H2O 3.36.1.2 documentation [web-document], URL: ⟨https://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html⟩ (accessed 7 January 22).
Videla, L.A., Fernández, V., Tapia, G., Varela, P., 2003. Oxidative stress-mediated hepatotoxicity of iron and copper: role of Kupffer cells. Biometals 16, 103–111. https://doi.org/10.1023/A:1020707811707.
Villeneuve, J.-P., Pichette, V., 2004. Cytochrome P450 and liver diseases. Curr. Drug Metab. 5, 273–282. https://doi.org/10.2174/1389200043335531.
Wang, Y., Zhang, S., Li, F., Zhou, Y., Zhang, Y., Wang, Z., Zhang, R., Zhu, J., Ren, Y., Tan, Y., Qin, C., Li, Y., Li, X., Chen, Y., Zhu, F., 2020. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res 48, D1031–D1041. https://doi.org/10.1093/nar/gkz981.
Whirl-Carrillo, M., McDonagh, E.M., Hebert, J.M., Gong, L., Sangkuhl, K., Thorn, C.F., Altman, R.B., Klein, T.E., 2012. Pharmacogenomics knowledge for personalized medicine. Clin. Pharmacol. Ther. 92, 414–417. https://doi.org/10.1038/clpt.2012.96.
Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M., 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.
Zhao, M., Zhang, T., Li, G., Qiu, F., Sun, Y., Zhao, L., 2017. Associations of CYP2C9 and CYP2A6 polymorphisms with the concentrations of valproate and its hepatotoxin metabolites and valproate-induced hepatotoxicity. Basic Clin. Pharmacol. Toxicol. 121, 138–143. https://doi.org/10.1111/bcpt.12776.