Challenges in Early-stage Prediction of Drug-Induced Liver Injury

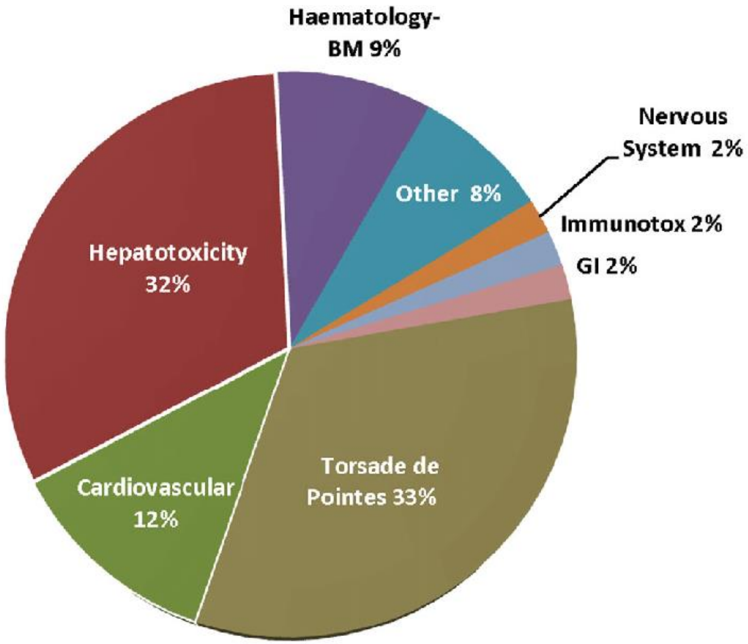
Jan Serroyen, Vahid Nassiri, Jeroen Tolboom, Kanaka Tatikola, Katja Waepenaert, Jan Snoeys

Nonclinical Biostatistics Conference
June 16, 2025, Rutgers University, New Brunswick

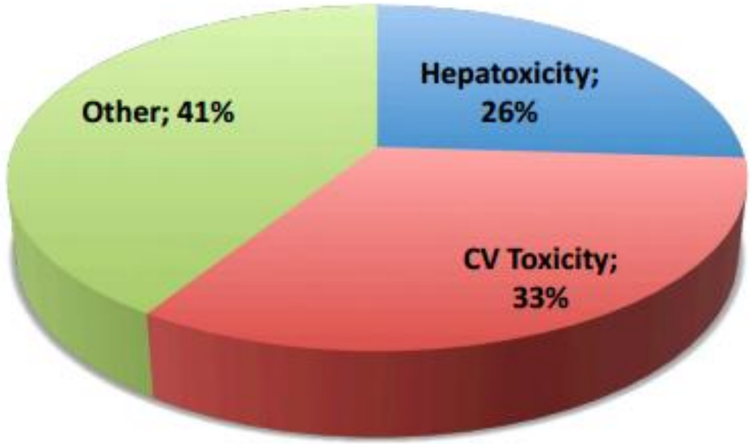
Johnson&Johnson
Innovative Medicine

Drug-Induced Liver Injury (DILI)

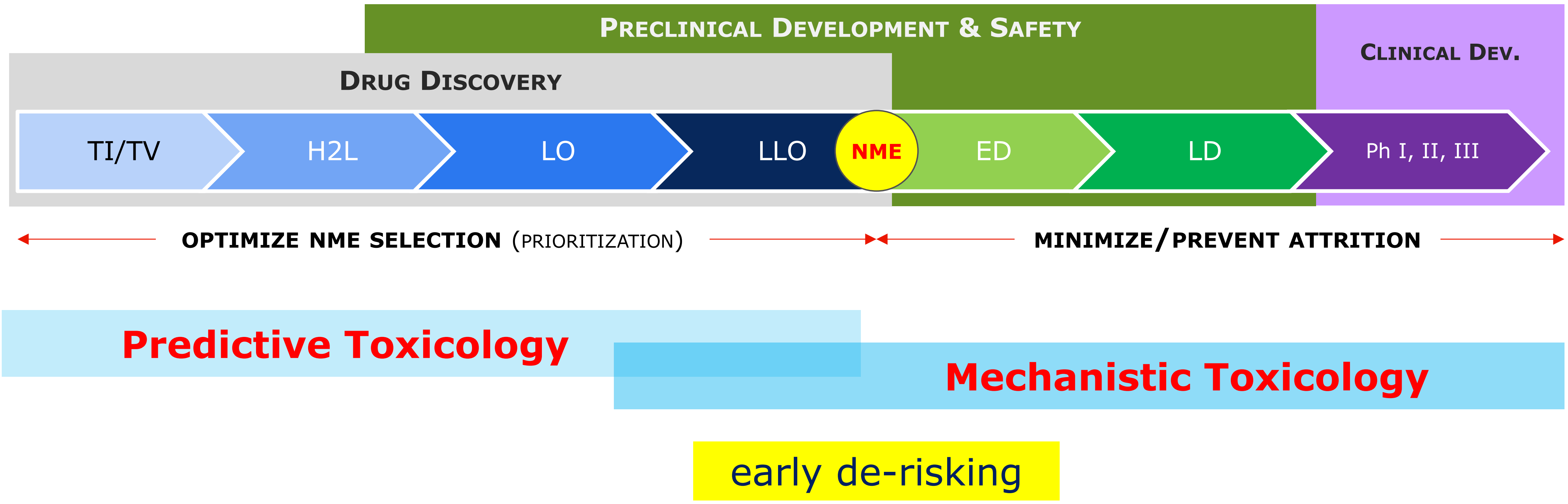
- Liver injury due to prescription and nonprescription medications
- DILI is a major concern for drug developers, regulators, and clinicians



Adverse drug events that have led to withdrawal from the marketplace worldwide between 1975-2007
[Stevens and Baker 2009]



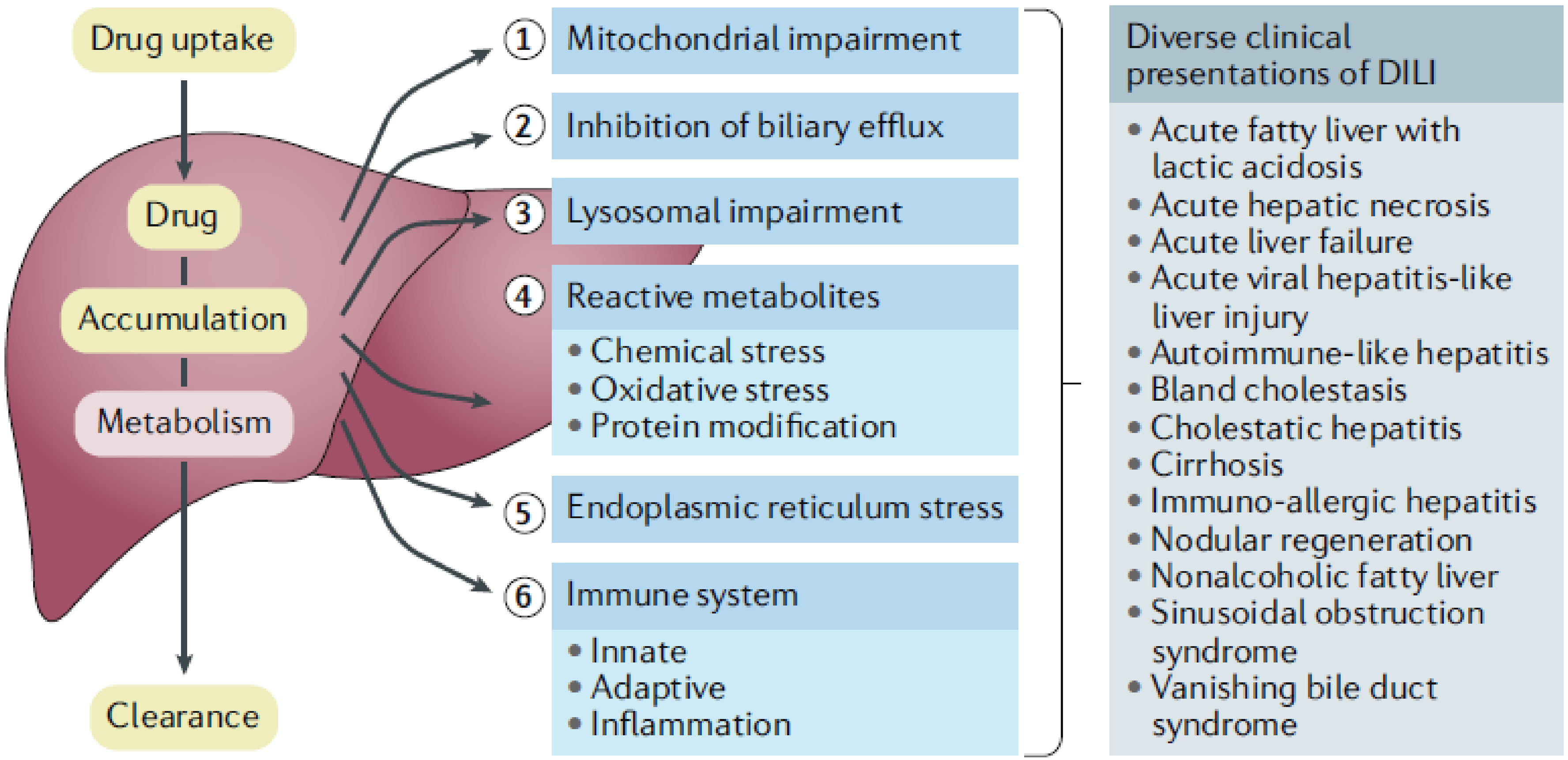
Drugs withdrawn from global market due to toxicity 1990-2010 (n=39)
From: EvaluatePharma; CDER; Tufts Center for Drug Discovery



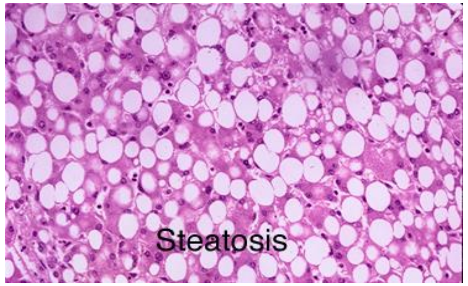
TI/TV: Target Identification / Target Validation
H2L: Hit-to-Lead
LO: Lead Optimization
LLO: Late Lead Optimization

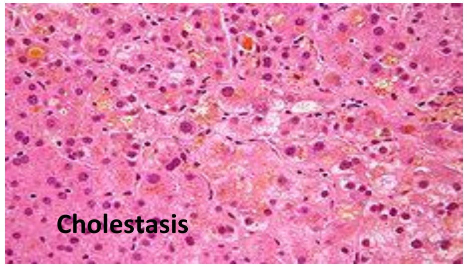
NME: New Molecular Entity
ED: Early Development
LD: Late Development

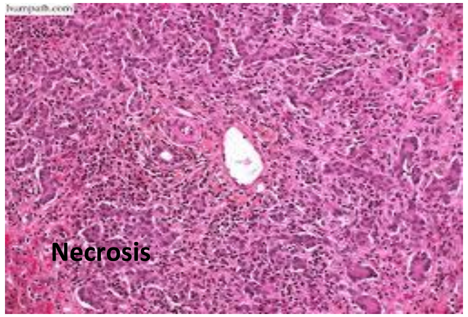
Drug-Induced Liver Injury (DILI)



DILI mechanisms **Clinical phenotypes**

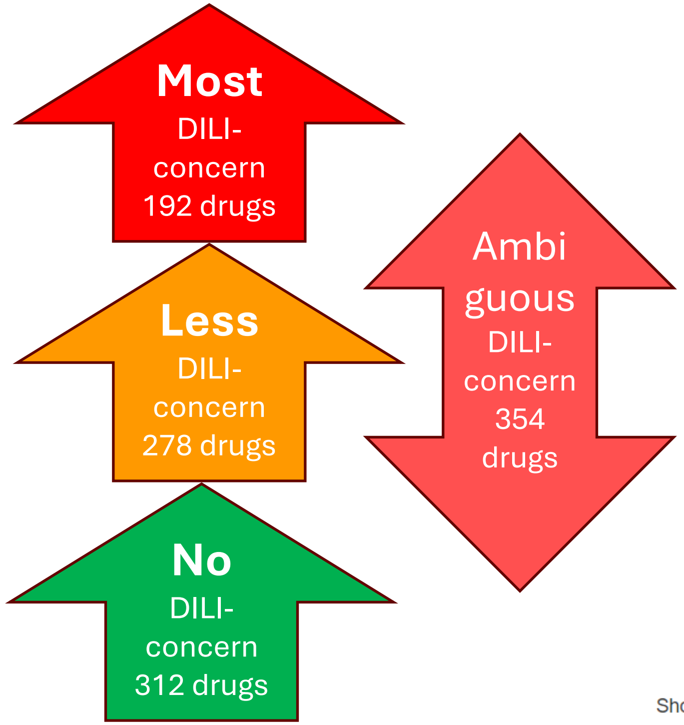






Weaver et al. (2020)

Categorizing DILI

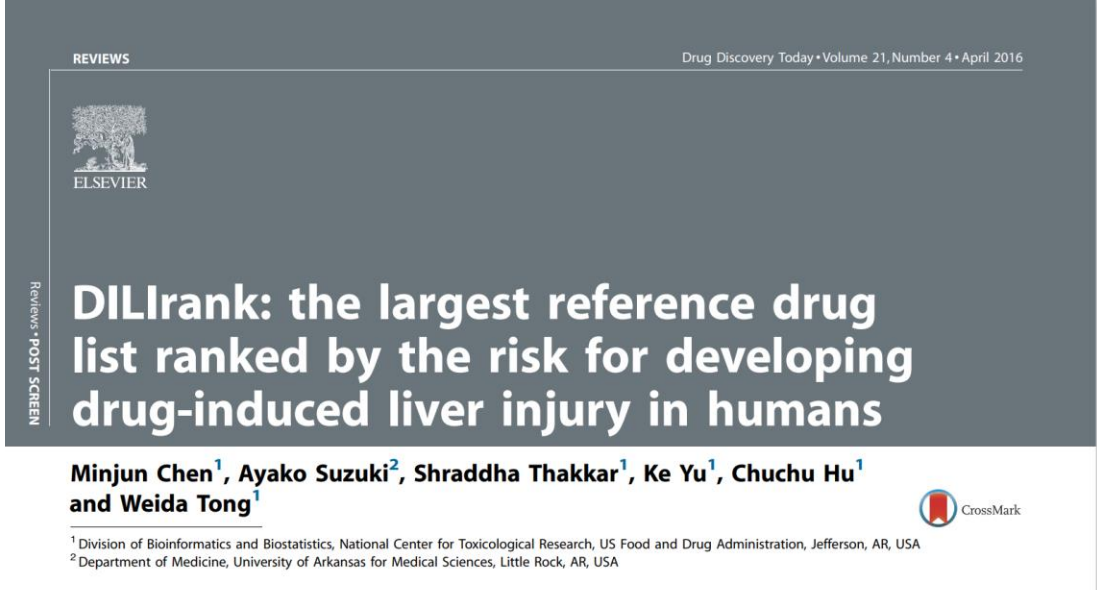


Search:

Show All entries

LTKBID	Compound Name	Severity Class	Label Section	vDILIConcern	Version
LT00003	mercaptopurine	8	Warnings and precautions	Most-DILI-Concern	1
LT00004	acetaminophen	5	Warnings and precautions	Most-DILI-Concern	2

[Drug Induced Liver Injury Rank (DILIrank) Dataset | FDA]


REVIEWS Drug Discovery Today • Volume 21, Number 4 • April 2016

ELSEVIER

Reviews • POST SCREEN

DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans

Minjun Chen[1], Ayako Suzuki[2], Shraddha Thakkar[1], Ke Yu[1], Chuchu Hu[1] and Weida Tong[1]

CrossMark

[1] Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA
[2] Department of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA

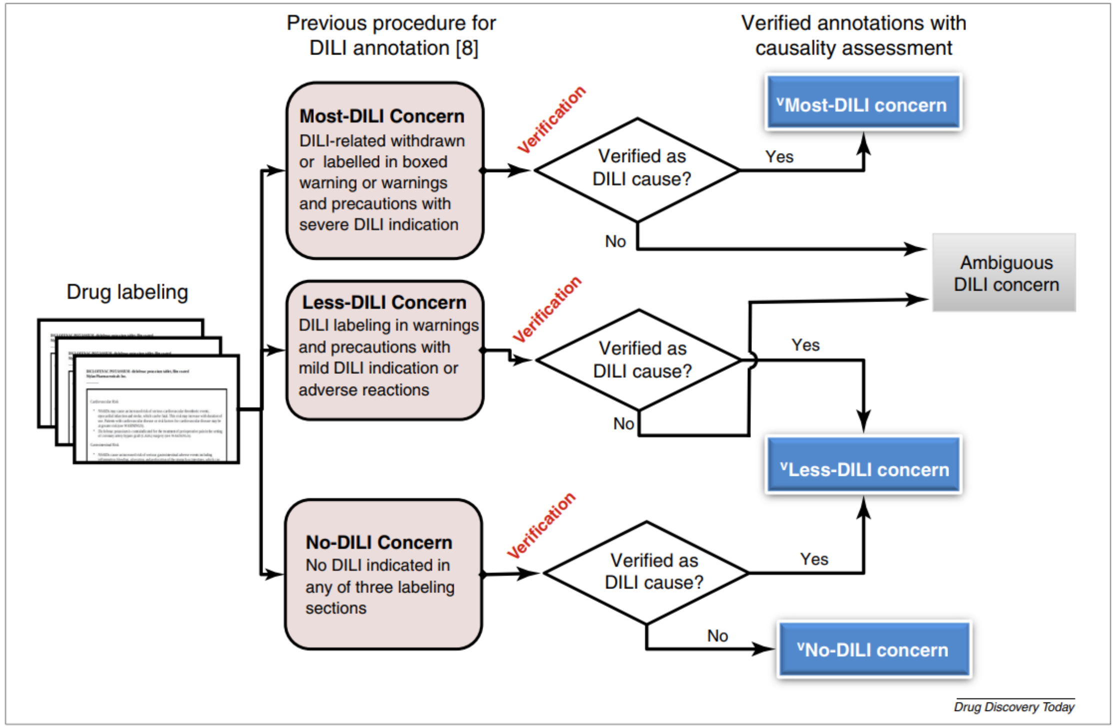
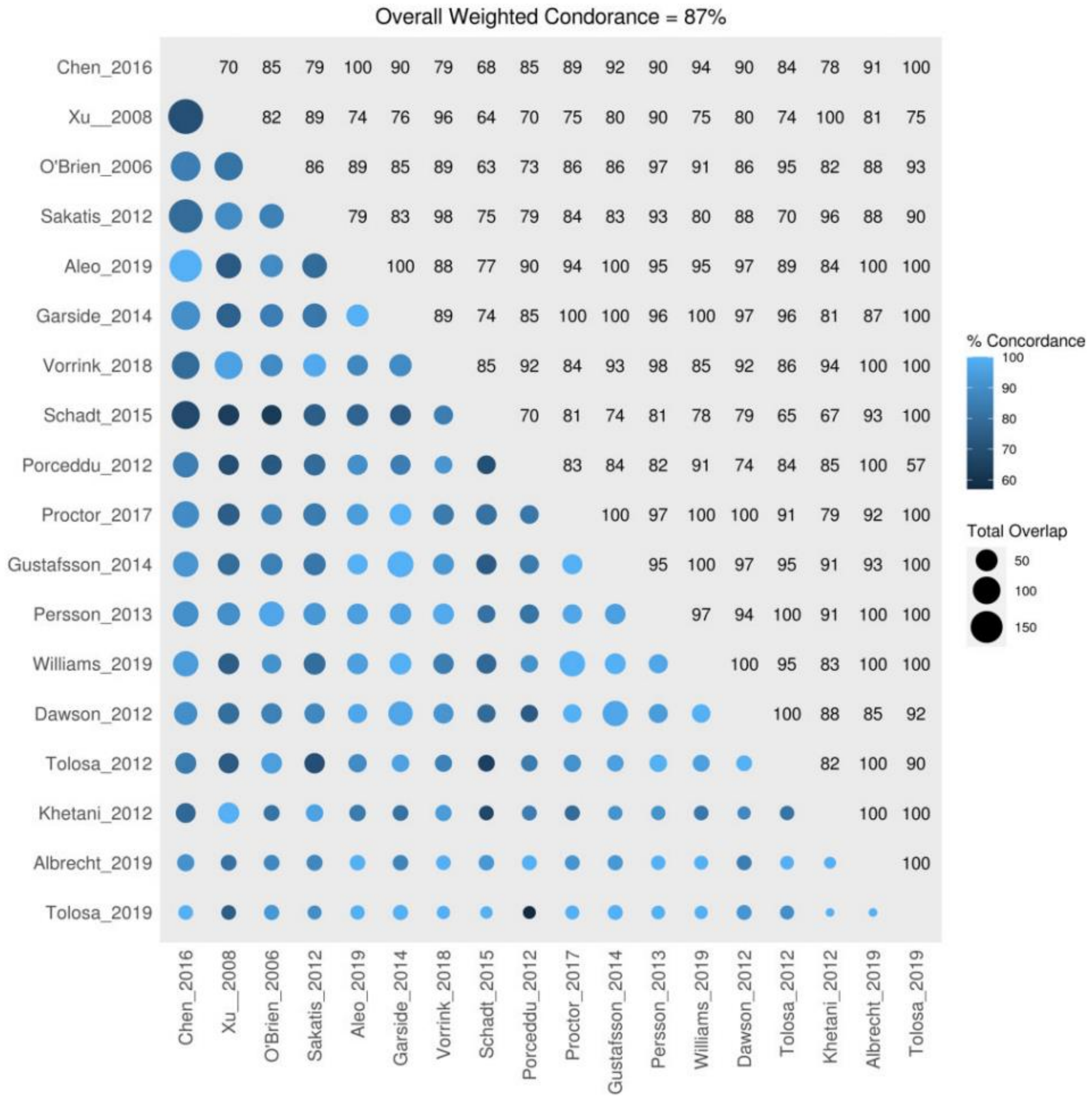


FIGURE 1

No Overall Agreement

- There is no universally accepted framework for categorizing drugs based on their potential to cause drug-induced liver injury (DILI).
- As an example, despite being placed in a Most DILI-concern category, acetaminophen is widely considered one of the safest and most commonly used over-the-counter medications **when taken at recommended dosages**.
- Liver-Tox book: LiverTox - NCBI Bookshelf (nih.gov)
  - Provides overview of hepatotoxicity for about 1000 drugs



Martin et al. (2022)

Predicting Drug-Induced Liver Injury

- Predicting drug-induced organ injury in early drug development is a **multi-dimensional** problem
- Requires multifaceted assays and compound-related information

In Vitro assays

- Cytotoxicity (IC50)
- Mitochondrial toxicity (HepG2 Glu/Gal ratio)
- Bile Salt Export Pump (BSEP) inhibition (IC50)

Physicochemical Properties

- Molecular weight (MW)
- Lipophilicity (cLogP)
- Spatial complexity (fsp^3)
- Ionization state (acid, base, or neutral)

Pharmacokinetic Properties

- Total plasma C_{max}
- Total daily dose
- Liver inlet concentration
- Fraction unbound

- Statistical approaches which **integrate data from different sources**
  → make the decision-making process faster, data driven, and more efficient

METHODS

Methods Evaluated within J&J

Single endpoint:

- Organoids / 3D cell (e.g. Fäs et al. 2025)

Multiple heterogeneous endpoints:

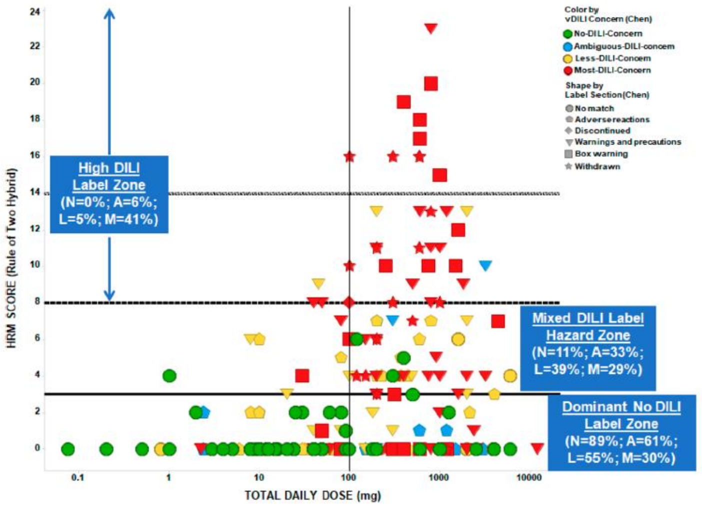
- In-vitro assays, physicochemical properties, pharmacokinetic properties (this talk)
- Omics: RNAseq, proteomics
- Off-target binding (e.g. Rao et al. 2023)
- Chemical structure (SMILES)
  - AI/ML methods (e.g. LLM)
  - In-silico predictions (e.g. Seal et al. 2024)

Multiparametric Approaches to DILI Prediction

Aleo et al. 2020

- **Decision rules**-based method that provides a quantitative DILI-risk
- Uses a scoring system based on safety margins (ICx/Cmax)
- Provides **cut-points** to separate M-DILI from N-DILI and L-DILI

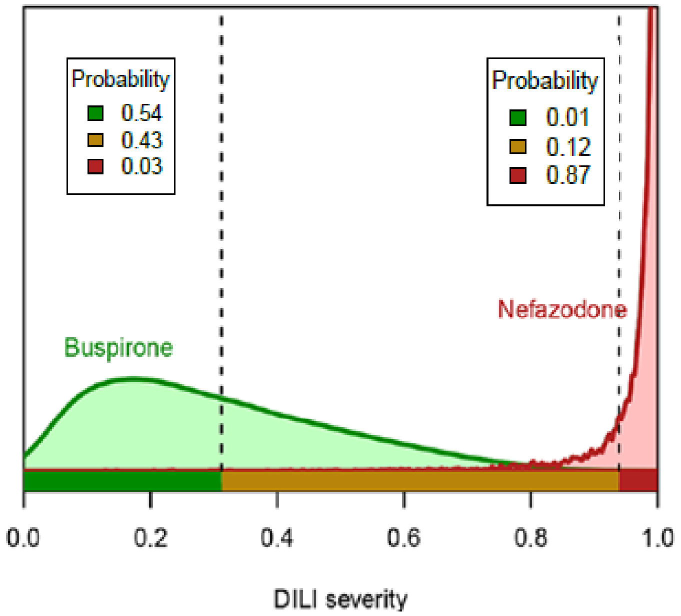
Scoring System



Williams et al. 2020

- Considers DILI severity-class as an ordinal variable
- **Bayesian statistical model** that provides a quantitative DILI risk
- Determines probabilities of each DILI severity class
- Separates three DILI severity classes from each other.

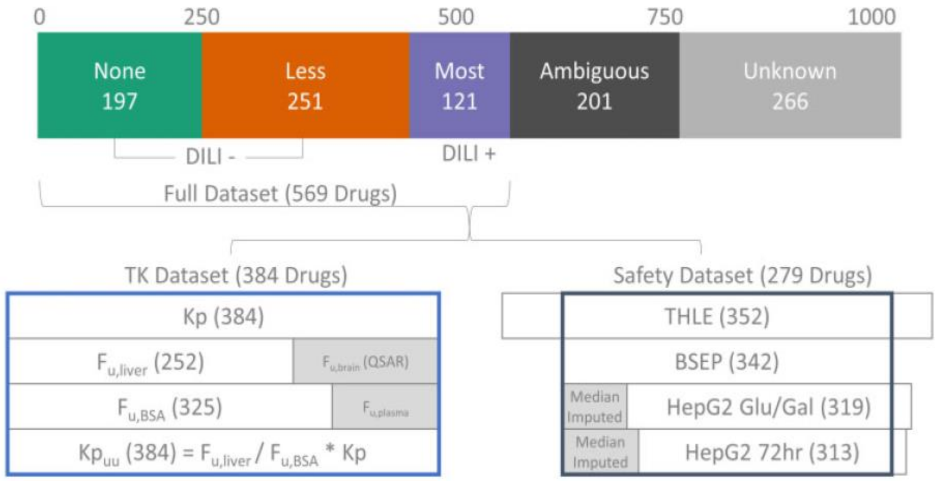
Bayesian Ordinal Logistic Regression (BOLR)



Martin et al. 2022

- Treats DILI as **binary**: N|L-DILI vs. M-DILI
- ML-based method that provides a quantitative DILI-risk
- Provides cut-point to separate M-DILI from the two other classes.

Random Forest



Scoring System (Aleo et al. 2020)

Scoring of Drugs Based on Activity in Core Mechanistic Assays

	Threshold ($IC_{50}/C_{max,total}$ value) Based Scoring					
Assay Category	≤1	>1 - ≤10	>10 - ≤50	>50 - ≤100	>100	Assay (Test Limits / %CV)[a]
Cytotoxicity (THLE or HepG2)	4	3	2	1	0	(300 µM / 27 or 29%)
Mitochondrial Dysfunction (inhibition)	4	3	2	1	0	(25 µM / 18%)
Mitochondrial Dysfunction (uncoupling)	4	3	2	1	0	(100 µM / 40%)
BSEP Inhibition	4	3	2	1	0	(100/200 µM / 34/48%)
	HepG2 Glu/Gal Ratio Scoring					
	≥3		≥2 - <3		<2	
Mitochondrial Dysfunction (cellular)	4		2		0	(300 µM / 21% glucose and 31% galactose)

[a]%CV = Percent coefficient of variation (interday) of positive controls (average).

Scoring of Drugs Based on Physicochemical Property Models

	Rule of Two Model Scoring Approach[21]
Rule of Two (total daily dose≥100 mg and cLogP≥3)	YES = 4, NO = 0
	Partition Model Scoring Approach[22]
Acids	cLogP≥2.5 = 4, OTHERWISE 0
Bases	-p[MDD][a] ≥ -3.49 and cLogP≥1.1 = 4, OTHERWISE 0
Neutrals	-p[MDD] ≥ -4.07 and Fsp3[b] <0.29 = 4 or -p[MDD] ≥ -4.07 and Fsp3 ≥0.29 and cLogP ≥1.0 = 4 OTHERWISE 0

Total Score: ≤3 = No-DILI
4-7 = Less-DILI
≥8 = Most-DILI

Example: Loratadine is an antihistamine and categorized as N-DILI by FDA.

	Cytotox		Mitotox				
Assay	HepG2	THLE	Glu/Gal	Inhibition	Uncoupler	BSEP	Cmax
Value	67.04	38.73	1.3	>25	>100	28.671	0.068
Normalized by Cmax	986.01	569.56	-	-	-	421.63	

Score = 0

Ionization	-PMDD	cLogP	Fsp3
Neutral	3.981	5.051	0.364

Score = 4 → **Total Score:** 0 + 4 = 4

J&J Innovative Medicine

Optimized Scoring System

Limitations of cut-points proposed by Aleo et al.

- Fixed scoring logic → cannot be adapted to specific objectives
- Specific to their dataset and assays → may not generalize well

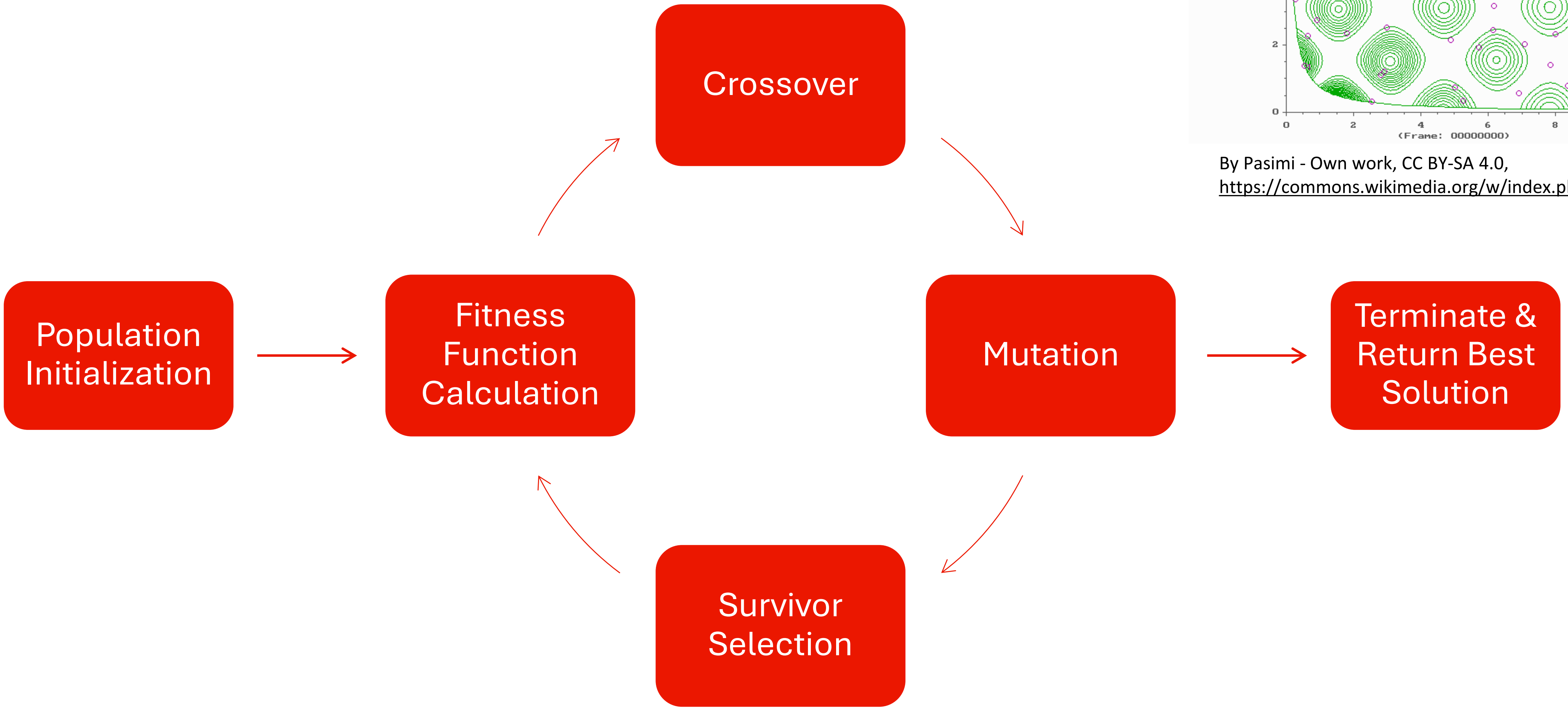
Proposal: Optimize cut-points

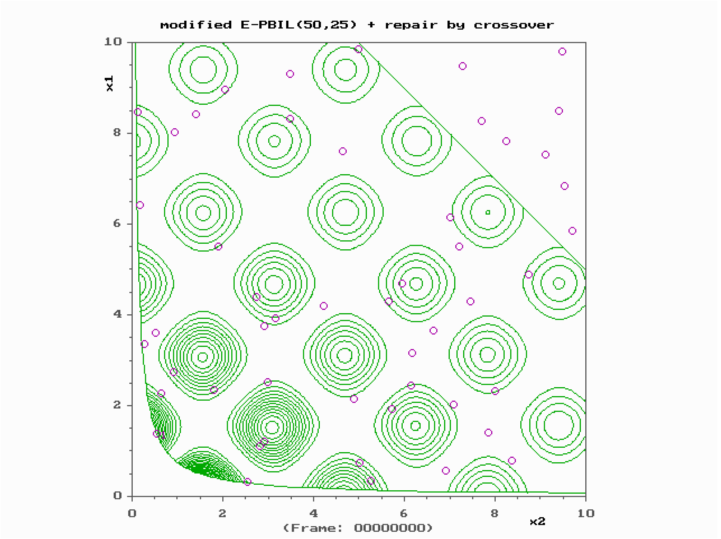
- Challenging optimization problem
  - ➢ discrete jumps at cut-points
  - ➢ complex relationship between endpoints
- Optimize for a specific objective
  - ➢ flexibility in objective function needed
  - ➢ example: maximize balanced accuracy

Solution: Genetic algorithm

Genetic Algorithm

Derivative free optimization algorithm





By Pasimi - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=37611586

Fitness Function

- Misclassifying Extreme cases to opposite class (i.e., No-DILI as Most-DILI & Most-DILI as No-DILI) should be avoided

  → Need for constraints: add penalty term to fitness function

- Goal: optimize **balanced accuracy**, while adhering to two **constraints**:
  - 80% of L/M-DILI correctly classified
  - 85% of N/L-DILI correctly classified

Note: false positive is considered worse than false negative (i.e. stopping a possible promising compound early)

Bayesian Ordinal Logistic Regression (Williams et al. 2020)

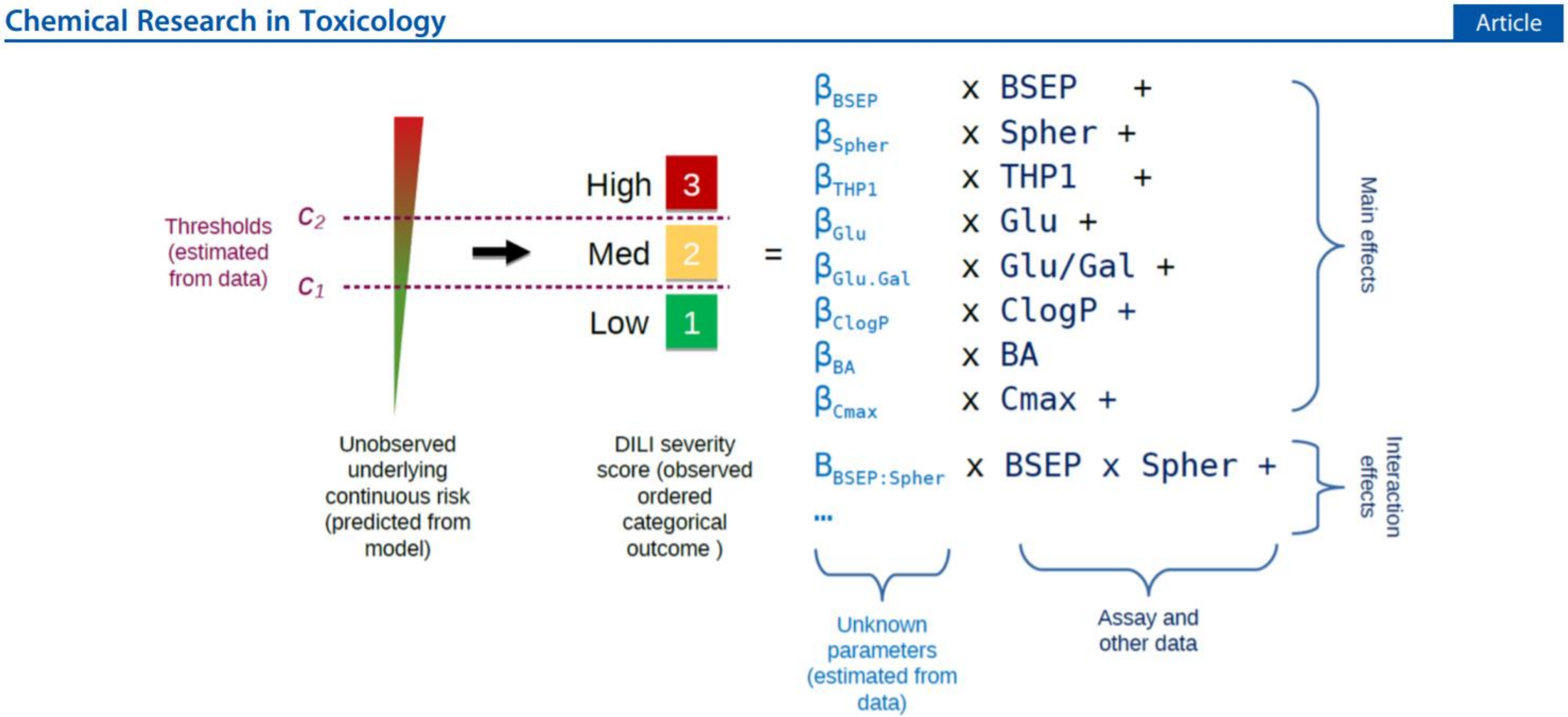


Figure 1. Bayesian predictive model. Clinically characterized DILI positive and negative compounds are classified according to their DILI severity score. Assays (BSEP, Spheroid, THP1, interaction effects, etc.) are used to estimate the unknown β values, which quantify the strength and direction of the relationship between the assay values and DILI severity, allowing prediction of the underlying continuous severity.

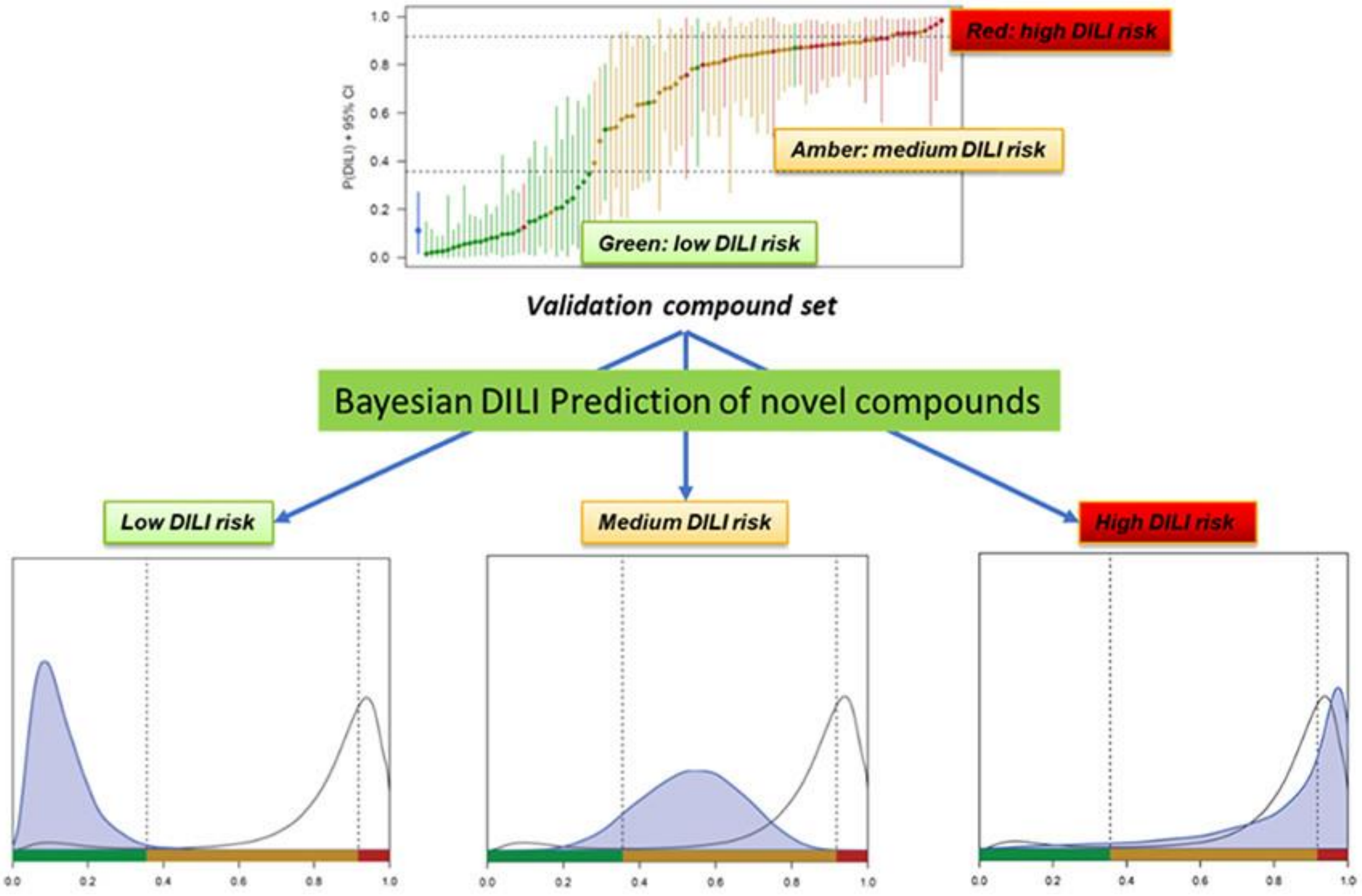
Bayesian Ordinal Logistic Regression (Williams et al. 2020)

Statistical model:

$$
\begin{aligned}
\text{Severity}_i &\sim \text{OrderedLogistic}(\eta_i, c_1, c_2) \\
\text{logit}(\eta_i) &= X_{ij} \times \beta_j \\
c_1, c_2 &\sim \text{Normal}(0, 20) \\
\beta_j &\sim \text{Laplace}(\mu, \sigma) \\
\mu &\sim \text{Normal}(0, 2) \\
\sigma &\sim \text{HalfNormal}(0, 0.5)
\end{aligned} \tag{1}
$$

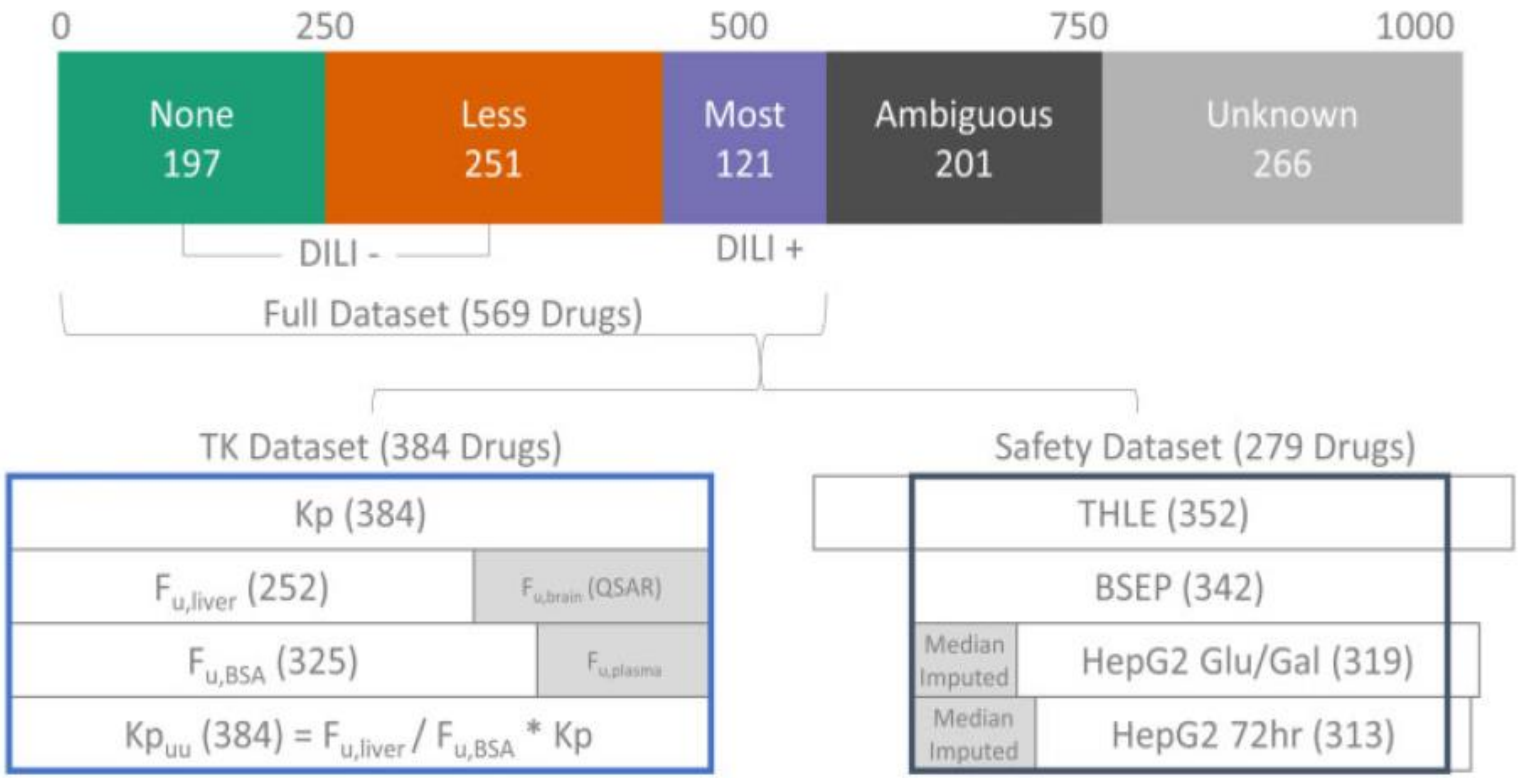
Here, η_i is the continuous prediction of DILI severity, and β_j are the regression coefficients. The model considers two-way interactions of all predictors except for log10.cmax. The model is fitted in `Stan` [2] via the `R` package `rstan` [6].

Bayesian Ordinal Logistic Regression (Williams et al. 2020)



Random Forest (Martin et al. 2022)

- Optimize ROC AUC
- Downsampling to deal with class imbalance



RESULTS

Performance Metrics

Actual DILI class

Predicted Class		Non-DILI	DILI
	Non-DILI	True Negative (TN)	False Negative (FN)
	DILI	False Positive (FP)	True Positive (TP)

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

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

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

Comparison of Original and Optimized Cut-points

Table 1: Cut-points from Aleo et al. (2019) and optimized cutpoints obtained using GA for multi-class DILI from Aleo et al. (2019) and binary DILI from Martin et al., 2022.

Cutpoint	Aleo et al.	Optimized (Aleo et al.)	Optimized (Martin et al.)
$cp_{\text{lipophilicity, acid}}$	2.50	0.07	3.08
$cp_{\text{lipophilicity, base}}$	1.10	3.70	3.25
$cp_{\text{lipophilicity, neutral}}$	1.00	0.20	3.81
cp_{fsp^3}	0.29	0.55	0.64
$cp_{-p[\text{MDD}],\text{base}}$	-3.49	-3.61	-3.26
$cp_{-p[\text{MDD}],\text{neutral}}$	-4.07	-3.82	-4.22
$cp_{\text{Glu/Gal},1}$	2.00	0.94	2.97
$cp_{\text{Glu/Gal},2}$	3.00	2.69	4.63
$cp_{\text{IC}_{50},1}$	1.00	0.61	4.06
$cp_{\text{IC}_{50},2}$	10.00	11.14	20.95
$cp_{\text{IC}_{50},3}$	50.00	14.75	37.55
$cp_{\text{IC}_{50},4}$	100.00	70.29	68.16
$cp_{\text{Score},1}$	4.00	3.15	-
$cp_{\text{Score},2}$	8.00	7.18	3.30

Cross-validated results (averaged over runs)

Performance

Table 2: Performance Metrics for Different Methods: Multi-class Aleo

Method	Accuracy			Misclassification		BA
	N-DILI	L-DILI	M-DILI	N as M-DILI	M as N-DILI	
RF	0.70	0.47	0.61	0.07	0.07	0.59
OSS	0.72	0.46	0.57	0.04	0.10	0.58
SS	0.76	0.45	0.50	0.00	0.17	0.57
RF + OSS (agree)	0.83	0.49	0.70	0.02	0.07	0.67
RF + SS (agree)	0.86	0.55	0.69	0.00	0.07	0.70

Table 3: Performance Metrics for Different Methods: binary Martin

Method	Sensitivity	Specificity	PPV	NPV	BA
RF	0.74	0.59	0.52	0.82	0.67
OSS	0.60	0.64	0.50	0.74	0.62
SS	0.14	0.97	0.71	0.66	0.55
RF + OSS (agree)	0.76	0.66	0.57	0.85	0.71
RF + SS (agree)	0.40	0.95	0.76	0.82	0.68

Agreement between methods increases reliability of predictions

Cross-validated results (averaged over runs)

Agreement Between Methods

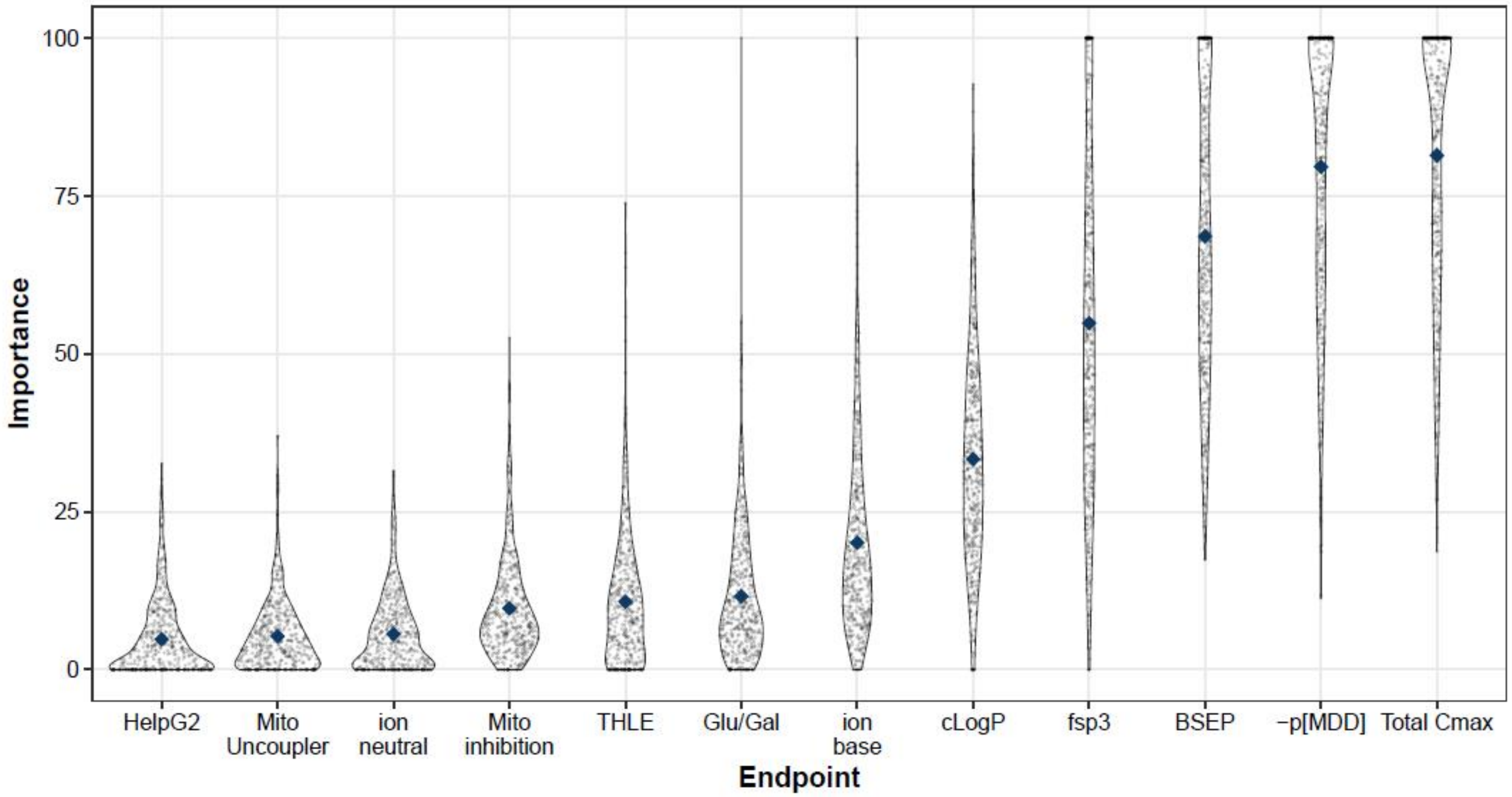
Table 4: Proportion of compounds for which the two methods agreed on the predicted class in Aleo et al.

	Method	N-DILI	L-DILI	M-DILI
1	RF + GA	0.71	0.54	0.63
2	RF + Aleo	0.69	0.42	0.58

Table 5: Proportion of compounds for which the two methods agreed on the predicted class in Martin et al.

	Method	N or L-DILI	M-DILI
1	RF + GA	0.69	0.67
2	RF + Aleo	0.60	0.37

Feature Importance in Random Forest Model



Results Optimized Scoring System

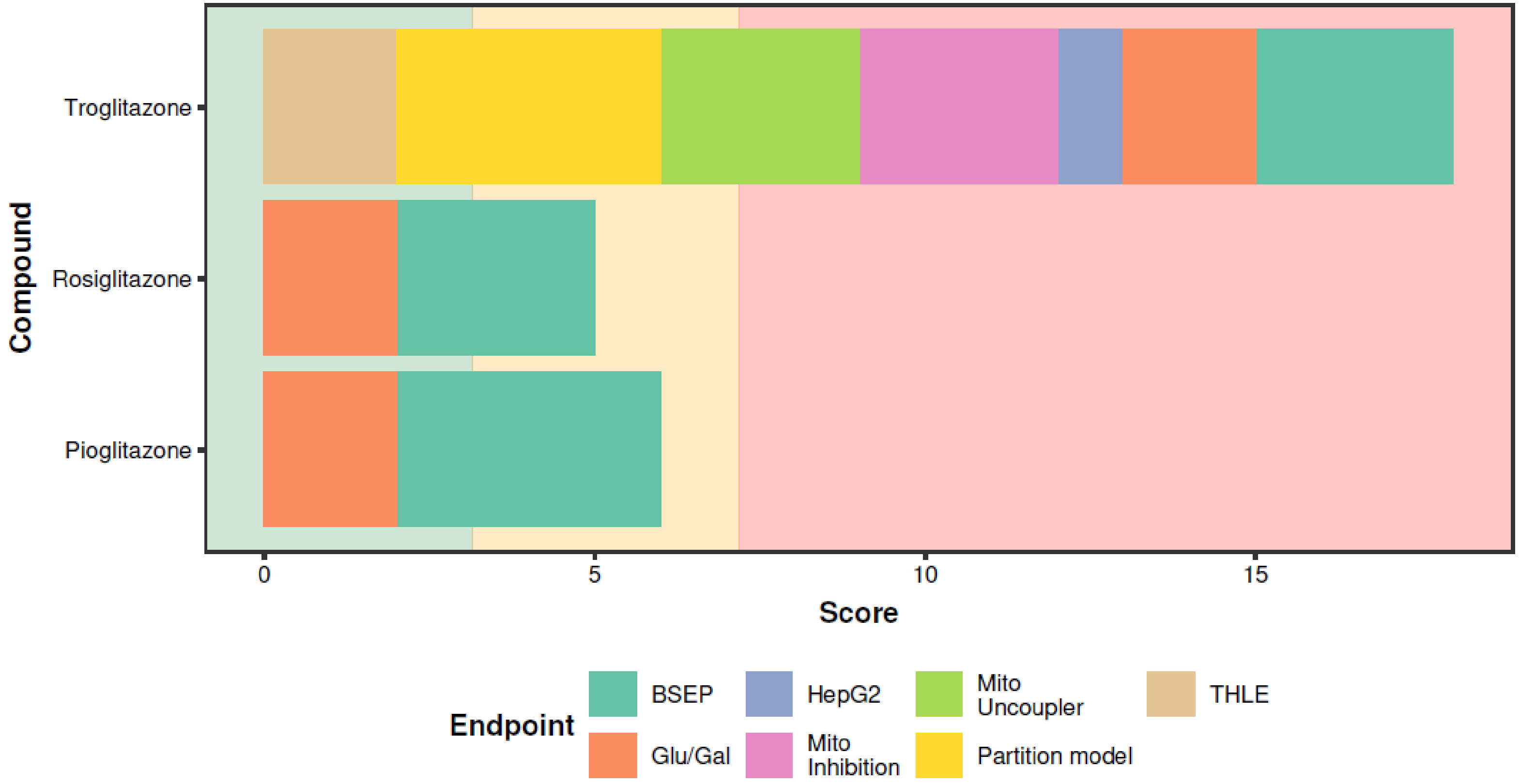
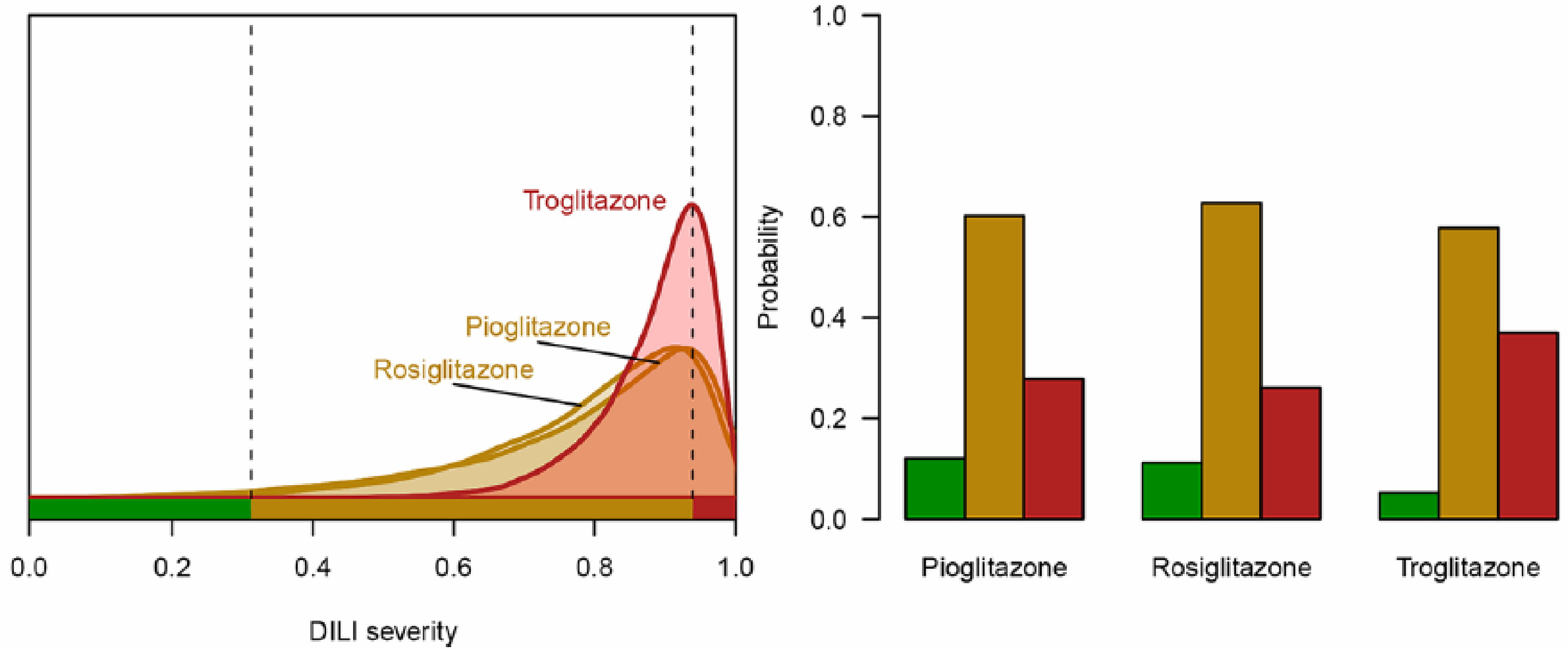


Figure 2: Computed DILI score for Pioglitazone, Rosiglitazone, and Troglitazone with the contribution of each endpoint to this score. The shaded areas show different predicted classes based on the optimized cutpoints.

Results Bayesian Ordinal Logistic Regression



(Williams et al. 2020)

Figure 4. Performance of the model on compounds with high molecular similarity and identical therapeutic indications, but different potential for hepatotoxicity. The order of severity of clinical hepatotoxicity is pioglitazone < rosiglitazone < troglitazone. Distribution color indicates the true DILI category. Graphs on the left show the posterior distributions (estimated continuous DILI severity), and bar graphs on the right show the predictions (posterior predictive distributions).

Take Home Messages

- ❑ DILI is a complex, multifactorial process
- ❑ Interpretability of predictions is a key requirement
- ❑ Optimized Scoring System combines expert knowledge with data-driven optimization
- ❑ Using different prediction methods and checking agreement increases robustness

References

- Aleo, M. D., Shah, F., Allen, S., et al. Moving beyond binary predictions of human drug-induced liver injury (DILI) toward contrasting relative risk potential. Chemical research in toxicology 33, 223–238 (2020).
- Chen, M., Suzuki, A., Thakkar, S., Yu, K., Hu, C., Tong, W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discov Today 21, 648–653 (2016).
- Fäs, L., Chen, M., Tong, W., et al. Physiological liver microtissue 384-well microplate system for preclinical hepatotoxicity assessment of therapeutic small molecule drugs. Toxicological Sciences, 203(1), pp.79-87 (2025).
- Martin, M. T., Koza-Taylor, P., Di, L., et al. Early Drug-Induced Liver Injury Risk Screening: “Free,” as Good as It Gets. Toxicological Sciences 188, 208–218 (2022).
- Rao, M., Nassiri, V., Alhambra, C., et al. AI/ML Models to Predict the Severity of Drug-Induced Liver Injury for Small Molecules. Chemical Research in Toxicology, 36(7), pp.1129-1139 (2023).
- Seal, S., Williams, D., Hosseini-Gerami, L., Mahale, M., Carpenter, A.E., Spjuth, O. and Bender, A. Improved detection of drug-induced liver injury by integrating predicted in vivo and in vitro data. Chemical Research in Toxicology, 37(8), pp.1290-1305 (2024).
- Stevens JL, Baker TK. The future of drug safety testing: expanding the view and narrowing the focus. Drug Discov Today, 14(3-4):162-7 (2009).
- Weaver, R.J., Blomme, E.A., Chadwick, A.E. et al. Managing the challenge of drug-induced liver injury: a roadmap for the development and deployment of preclinical predictive models. Nat Rev Drug Discov 19, 131–148 (2020).
- Williams, D. P., Lazic, S. E., Foster, A. J., Semenova, E., Morgan, P. Predicting drug-induced liver injury with Bayesian machine learning. Chemical research in toxicology 33, 239–248 (2020).

Thank You!

Johnson&Johnson
Innovative Medicine