cencora PharmaLex

janssen PHARMACEUTICAL COMPANIES OF Johnson&Johnson

Drug-Induced Liver Injury: Integrating Gene Expression Data and Target-Based Drug Families into an Advanced Predictive Toxicity Model

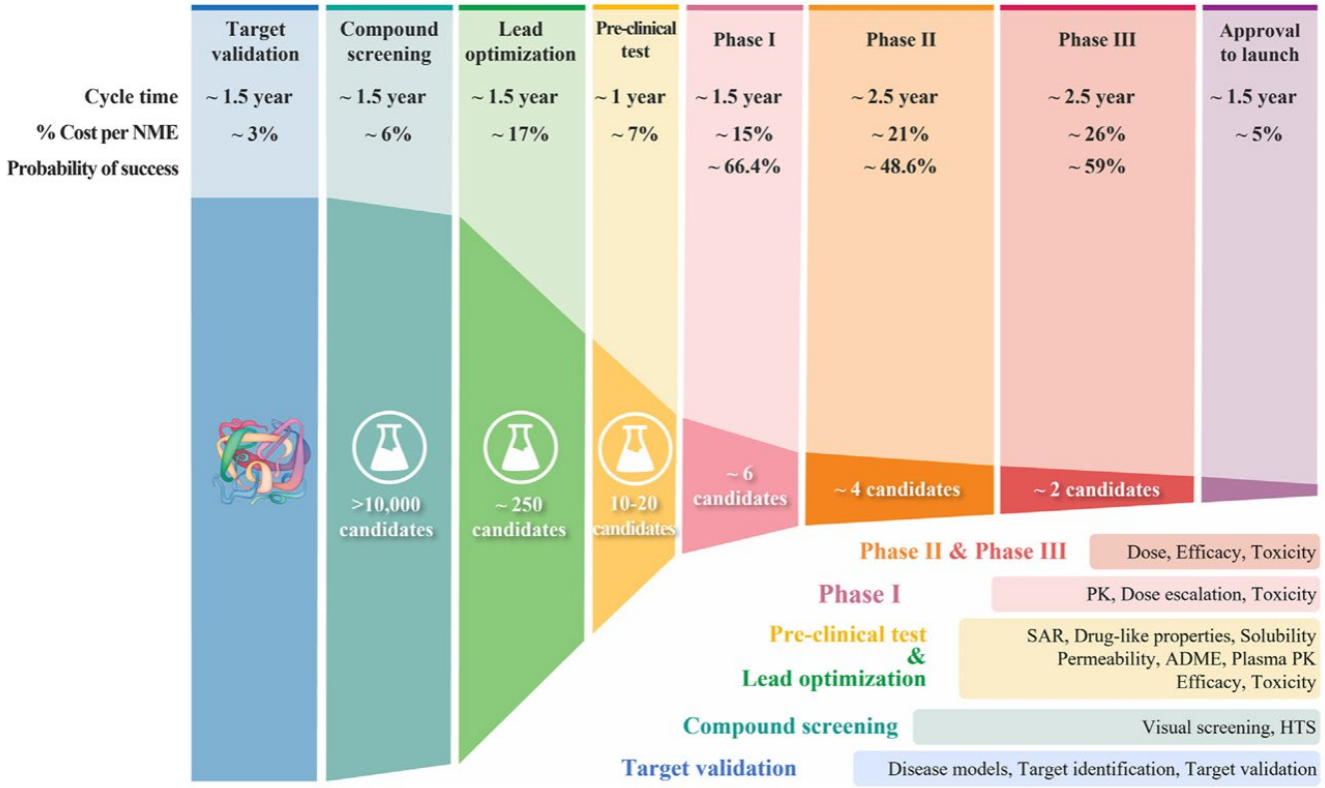
NCS 2024

Thomas de Marchin, Shreyas Hirway, Fetene Tekle, Xian Yao and Lira Pi

Outline

1. Introduction
    - Drug-Induced Liver Injury (DILI): A Challenge in Drug Development
    - Machine Learning for DILI prediction
2. Data Characteristics and Challenges
3. Method and Results
    - Machine Learning Pipeline
    - Models
    - Random effects in ML
    - High dimensional data: feature reduction (DE) and feature transformation (pathway enrichment)
    - Tested algorithms: Elastic Net and Tree-based methods
4. Conclusion and Discussion

Drug development pipeline



Toxicity assessment is a critical step in drug development to ensure safety

Sun, Duxin, et al. "Why 90% of clinical drug development fails and how to improve it?." *Acta Pharmaceutica Sinica B* 12.7 (2022): 3049-3062.

Drug-Induced Liver Injury (DILI): A Challenge in Drug Development

The Liver: One of the most susceptible organs to drug toxicity.

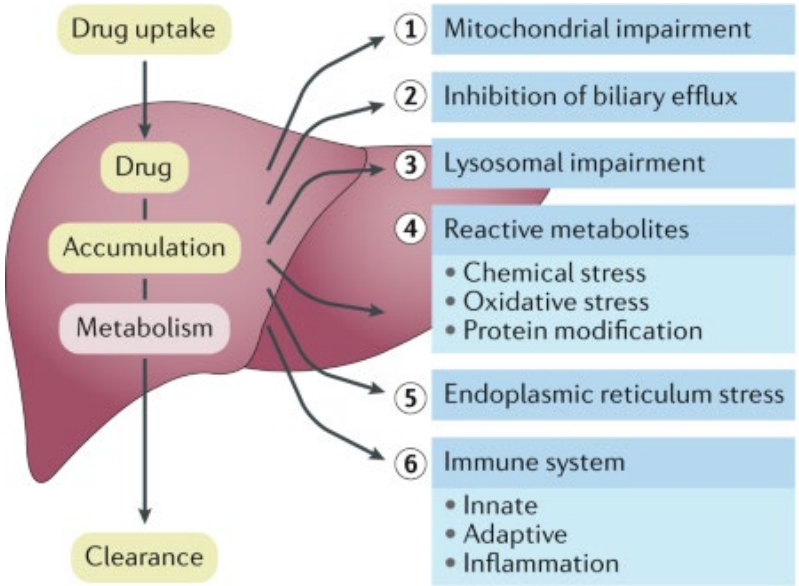
Drug-Induced Liver Injury (DILI): Leading cause of drug withdrawals and clinical trial failures.

DILIrank dataset:

- FDA resource with 1,036 drugs categorized by DILI risk:
  - Most Concern
  - Less Concern
  - No Concern
- Data from drug labels, literature, and real-world evidence

DILI Prediction

- **Why:** Traditional methods rely on animal testing (time-consuming, expensive, ethical concerns).
- **Challenge:** DILI in humans arises from diverse metabolic pathways and population heterogeneity (sex, age, genetics…), making it difficult to predict.



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).

Machine Learning for DILI Prediction: Challenges and New Directions

Multiple ML Models Explored (SVM, RF, kNN, EL, NN): Used to predict DILI, mainly from chemical structures or gene expression.

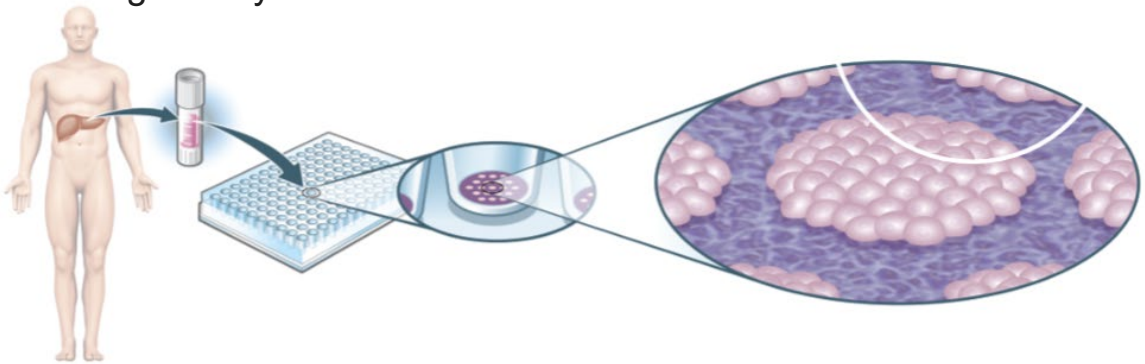
Gene Expression Models: Promising, but Limited

Achieved high accuracy but based on rat data (not applicable to humans)

Exception: Chierici et al. (2020) reported poor performance (random labelling) with human data, suggesting limitations of previous approaches

Towards in vitro Human-Relevant Models

Leverages human liver cell data (transcriptomics) from HEPATOPAC system: stable in vitro model for long-term drug toxicity studies.



https://bioivt.com/hepatopac-cultures

The transcriptomic data

18574 Genes

171 drugs

Drug	DILI	Dosage.uM	Cluster	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gen
Drug 1	1	1.00	B	-0.2212	0.1864	-0.0167	0.5074	0.1061	0.05	0.5368	0.28
Drug 1	1	0.11	B	-0.3239	0.2235	0.0656	0.7202	0.1539	0.3229	0.1511	0.13
Drug 1	1	0.33	B	-0.3472	-0.0003	0.216	0.4247	-0.2276	-0.1667	0.2716	0.06
Drug 2	1	1.00	A	-0.1236	0.3284	0.2368	0.3964	-0.2236	0.3151	0.8603	0.12
Drug 2	1	0.11	A	-0.0918	-0.00015	-0.001	-0.15485	0.3353	0.00185	0.207	0.24
Drug 2	1	0.04	A	-0.2697	-0.1985	0.0408	0.0344	-0.1663	-0.2057	0.317	0.39
Drug 2	1	0.33	A	-0.07585	-0.0445	-0.03195	0.15395	-0.11125	0.2403	-0.387	0.31
Drug 3	0	1.00	B	-0.2021	0.2945	-0.0469	-0.8307	0.0868	-0.2924	0.1175	0.05
Drug 3	0	0.14	B	-0.3751	0.2728	-0.0461	-0.2345	0.2554	-0.0952	0.0761	-0.1
Drug 3	0	0.38	B	-0.4285	0.4749	-0.0568	-0.9056	-0.0466	-0.0497	-0.2072	0.06
Drug 4	0	0.40	C	0.477	-0.076	0.1546	0.0952	-0.0837	-0.2125	-0.1064	-0.04
Drug 4	0	1.00	C	0.4699	-0.1618	0.2309	0.1315	0.0546	0.1757	0.1145	-0.1
Drug 5	1	1.00	C	-0.3106	0.3193	0.1307	0.5802	-0.3402	-0.0391	0.263	0.25
Drug 5	1	0.38	C	-0.3083	-0.3522	0.2114	-0.2592	-0.0966	0.0124	0.1944	0.09
Drug 6	1	1.00	B	-0.349	0.4417	-0.0243	-0.1663	0.11	-0.1771	0.1187	-0.05
Drug 6	1	0.14	B	-0.41	0.1021	0.0842	-0.2402	-0.0594	-0.1506	0.0842	0.21
Drug 6	1	0.38	B	-0.1607	0.3318	0.1113	0.0418	0.0615	-0.1117	-0.192	-0.03
Drug 7	1	1.00	A	-0.1251	-0.3291	-0.0538	-0.0841	-0.2413	-0.1658	-0.0964	0.16
Drug 7	1	0.16	A	0.1825	0.2352	-0.2181	-0.2552	-0.4156	-0.2614	-0.3477	0.4
Drug 7	1	0.40	A	0.0292	-0.0269	-0.4174	0.2545	0.1458	0.1873	0.57	0.23
Drug 8	1	1.00	A	0.0419	-0.0039	-0.2897	-0.1065	-0.0751	-0.1662	-0.181	0.23
Drug 8	1	0.16	A	0.32375	-0.2777	-0.25695	-0.6081	-0.0714	-0.2366	-0.38875	0.16
Drug 8	1	0.40	A	0.0533	-0.311	0.2318	-0.5949	-0.422	-0.1776	-0.5021	0.12
Drug 9	1	1.00	A	0.165	0.0873	-0.1122	0.3768	-0.044	0.1672	-0.4079	0.04
Drug 9	1	0.16	A	-0.1282	0.2009	0.017	-0.3499	-0.0348	0.2599	0.1	0.04
Drug 9	1	0.40	A	-0.124	-0.264	-0.1261	0.601	0.1208	0.5867	-0.3119	0.19
Drug 10	1	1	A	0.1488	-0.0913	0.0716	0.0794	-0.0955	-0.0615	-0.3769	-0.15
Drug 10	1	0.16	A	0.3212	0.2559	-0.2566	-0.1851	-0.1644	0.1668	-0.0697	-0.13
Drug 10	1	0.4	A	0.0361	0.0018	-0.2235	0.5863	0.0423	0.0043	-0.2316	-0.00
Drug 11	1	0.125	C	0.1797	0.1371	0.6857	0.2271	0.2009	0.0724	0.1826	0.14

Objectives

Primary Objective:

Predict Drug-Induced Liver Injury (DILI) from human hepatocyte gene expression data provided by Janssen.

Secondary Objective:

Investigate if incorporating drug family information (based on off-target genes, next slide) improves the model's ability to predict DILI. This would support the idea there are different drug families with distinct gene signatures for DILI.

Objective 2 – Drug family clustering by off-target genes

The Hypothesis:

- Traditional approaches: all drugs have the same DILI mechanisms.
- Our hypothesis: different drug families might have unique DILI toxicity pathways.

Drug family clustering algorithm:

- Off-target genes (genes unintentionally affected by drugs) were extracted from external databases (OTSA, ABCD, Off-X, DrugBank).
- Using data reduction techniques and clustering algorithms, we grouped the drugs into 3 distinct "families"

Testing the Theory

Two predictive models: without or with drug-clusters effect. Were the predictive performance improved ?

Table 1:
Number of coumpounds per cluster.

DILI_Binary	**N**	**0**, N = 42[1]	**1**, N = 145[1]
PCAKmeans3_250_15	173		
A		9 (24%)	55 (41%)
B		22 (58%)	47 (35%)
C		7 (18%)	33 (24%)
Unknown		4	10

[1] n (%)

Data Characteristics and Challenges

Limited samples: 171 Drugs available for training and testing

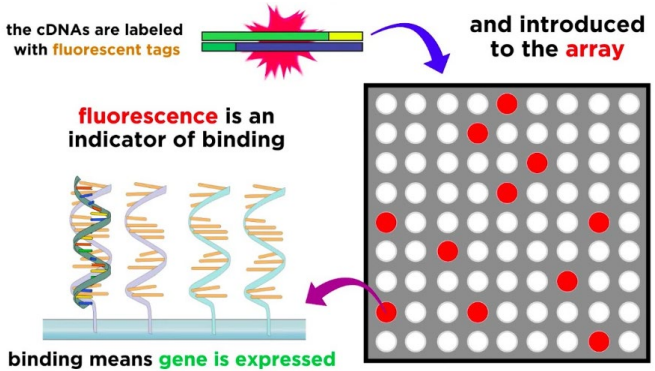
Imbalanced Classes: 80% DILI 1 and 20% DILI 0

18574 genes (features) measured using microarray

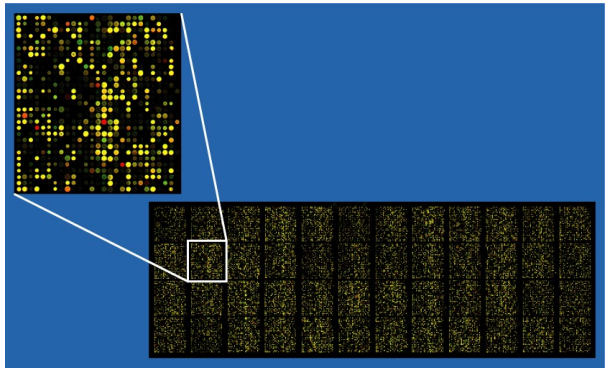
Single Donor Data: using a single donor simplifies experimentation (high-throughput testing of drug candidates) but may not capture the full spectrum of human responses:

- Diversity: Genetic & mitochondrial differences among people can affect how they metabolize drugs
- DILI can be a rare event (estimated incidence of 1 per 100,000 treated patients for some drugs): A hepatotoxic drug might not be toxic for our donor

Multiple Doses: identifying the most informative dose for prediction remains unclear (effective vs. too high).

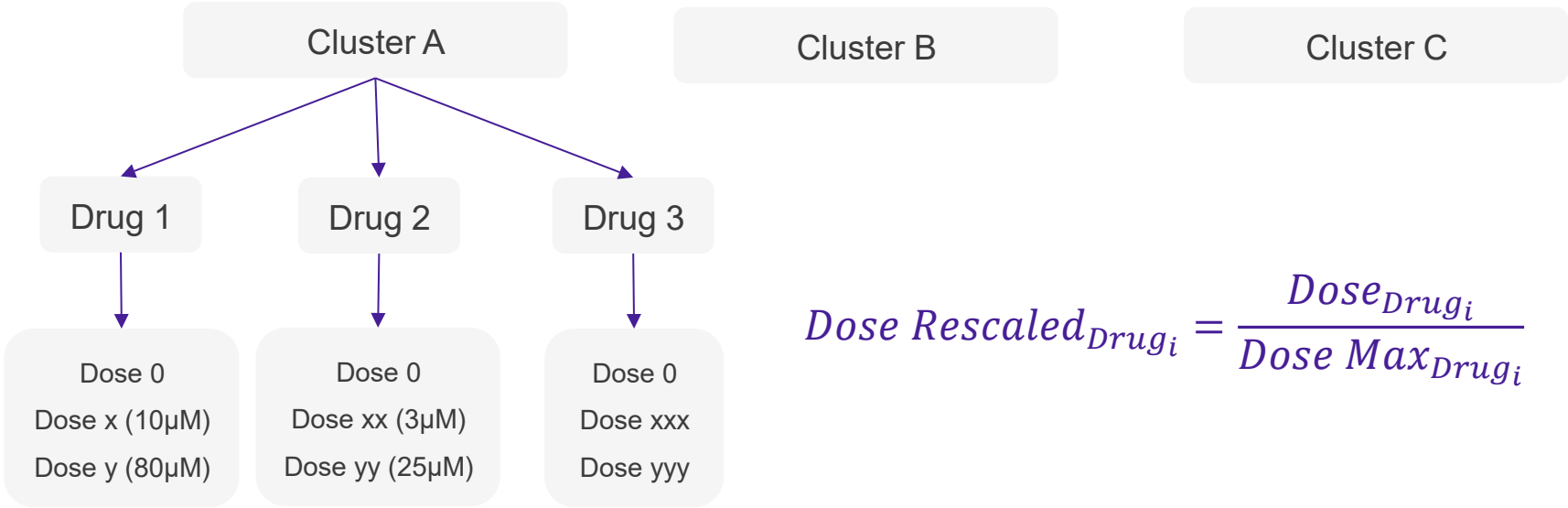


https://www.youtube.com/watch?app=desktop&v=Hv5flUOsE0s



https://en.m.wikipedia.org/wiki/File:Microarray2.gif

Dose-level calibration



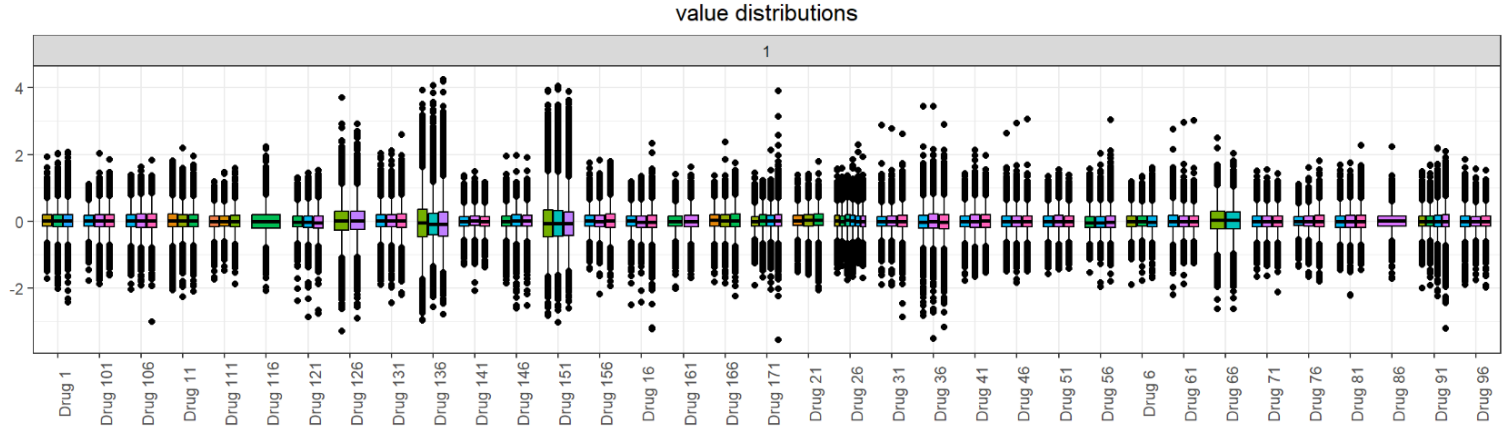
$$Dose\ Rescaled_{Drug_i} = \frac{Dose_{Drug_i}}{Dose\ Max_{Drug_i}}$$

Inter-Drug Variability: Direct comparison of doses across different drugs is challenging

- Drugs have varying potencies and mechanisms of action
- 100 μM of Drug A does not have the same effect as 100 μM of Drug B

Rescaling Approach: the doses were rescaled to a range of 0 to 1, by drug

Distribution of normalized transcriptomic data



Gene Expression Data Preprocessing

Data were normalized by subtracting the median value of control (concentration 0 / DMSO), per probe.

Plot

Distribution of normalized expression values for a subset of the 171 drugs tested. Each drug was tested at multiple concentrations (1-5). Each boxplot summarizes the expression data of 18,754 genes.

Machine Learning Pipeline

Algorithms:

- Elastic net penalized regression (*glmnet*) : *tidymodels*
- Random forest *(ranger): tidymodels*
- Gradient boosting (*LightGBM*): *tidymodels*
- Mixed effect gradient boosting (*GPBoost*): package functions

Data splitting:

- 75%/25% Grouped Stratified train/test split, repeated 10 times
- Hyper parameter optimisation: Grouped Stratified 5-Fold Cross-Validation, optimized on ROC–AUC
- Grouped splitting: observations from the same compound are never present in both training and testing data, preventing bias.

High Dimensionality: Feature reduction or feature transformation

Imbalanced Classes (more DILI cases): data upsampling (SMOTE) or class weighting

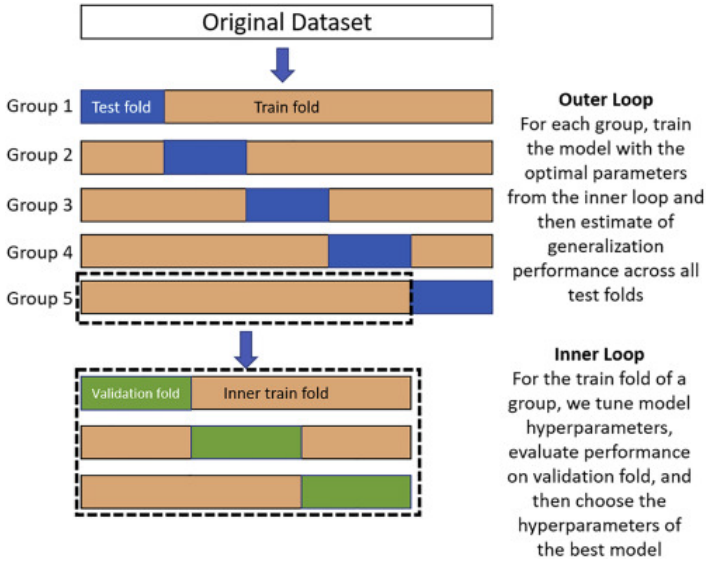


Image: Gupta, Abhijit, Mandar Kulkarni, and Arnab Mukherjee. "Accurate prediction of B-form/A-form DNA conformation propensity from primary sequence: A machine learning and free energy handshake." Patterns 2, no. 9 (2021).

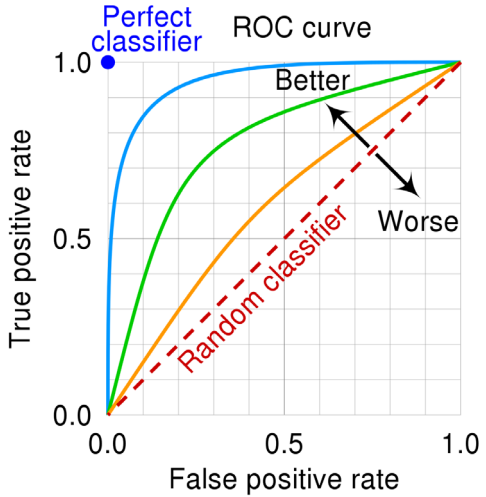
Evaluating Model Performance

ROC-AUC:

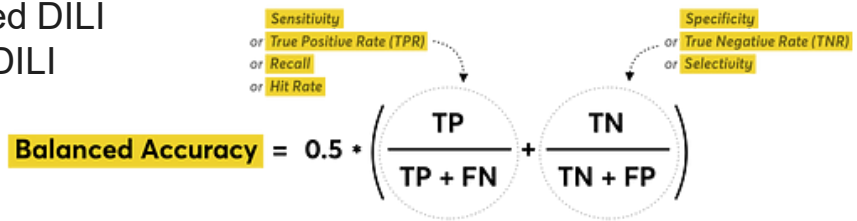
- Considers all possible classification thresholds for the model's predictions.
- AUC (Area Under the Curve): how well the model distinguishes between DILI and non-DILI cases across these thresholds.
- A score above 0.5 indicates some predictive power

Balanced Accuracy:

- Imbalanced data (unequal numbers of DILI and non-DILI cases): provides a more reliable assessment than accuracy.
- arithmetic mean of sensitivity (correctly identified DILI cases) and specificity (correctly identified non-DILI cases).



https://medium.com/@ilyurek/roc-curve-and-auc-evaluating-model-performance-c2178008b02



https://towardsdatascience.com/performance-metrics-for-binary-classifier-in-simple-words-be958535db49

Filtering: Dimension reduction using Differential Expression analysis

$$Gene_{ij(k)l} \sim \beta_1 * DILI_l + \beta_2 * Dose + (\beta_3 * Cluster_k) + \gamma_1 * Drug_l + \varepsilon$$

With

- $Gene_{ij(k)l}$: the expression level of a particular gene (i) in a sample (j) for the drug (l).
- $DILI_l$ (0 or 1): the DILI value for the l-th drug
- $Dose$ (0 to 1): the normalized dose of the drug in the sample
- Optional - $Cluster$: the cluster assigned to the drug (A, B or C)
- $Drug_l$: the l-th drug, random effect assumed normally distributed
- ε: error term

Challenge: Our data includes a massive number of genes (18,574). Analyzing them all is computationally complex and would lead to a poor model.

Feature Selection with LIMMA: popular R package for transcriptomic data. Widely used for the detection of differentially expressed genes between experimental conditions.

Identifying the Most Promising Genes:

- Calculate a score (p-value) for each gene (contrast DILI1 vs DILI0).
- Low p-values: potential association with DILI.
- Kept the top 50 most significant genes for inclusion in the predictive model

Accounting for Drug-Specific Effects with Random Effects

Challenge: Correlated observations

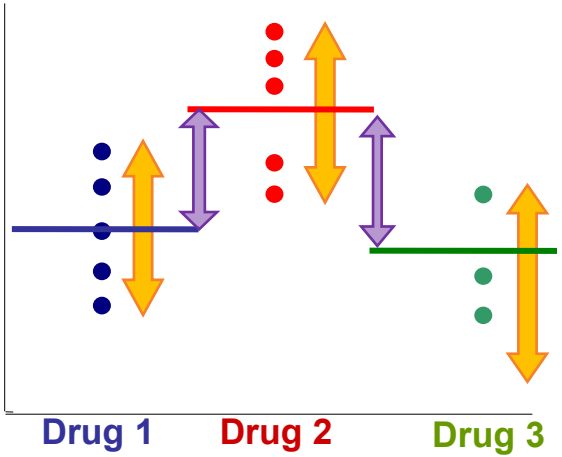
Doses from the same drug: naturally correlated → each drug has its own unique characteristics beyond the genes associated with DILI.

Random Effects: Capturing Drug Variability

- While random effects models are common in statistics, their application in bioinformatics seems less widespread
- Unlike fixed effects (focusing on individual drugs), random effects capture the overall influence of drug-to-drug variations: we're not interested in the specific effect of each individual drug

Why it Matters

- Most (if not all) ML algorithms assume that training samples are independent → often violated (repeated observations, longitudinal data).
- Not modeling samples correlation can lead to mediocre performances or potential misleading inferences.



Model(s)

$$DILI_l \sim f(Genes + Dose + (Cluster)) + \gamma\, Drug_l$$

With:

- $DILI_l$ (0 or 1): the DILI value of the l-th drug
- $Genes$: the expression values of the microarray probes
- $Dose$ (0 to 1): the normalized drug dose of each sample
- Optional - $Cluster$: the cluster assigned to a drug (A, B or C)
- $Drug_l$: the l-th drug, random effect assumed normally distributed

Correlated Data: Measurements from the same drug are correlated. Including drug as a fixed effect would treat each drug uniquely, making predictions for unseen drugs impossible → we will fit a mixed model with drug as a random effect.

Tree-based ML algorithms with Random Effects

$$y \sim F(X) + Zb + \varepsilon$$

With:

- $F(X)$: trees based on the X "fixed" predictors (features)
- Zb: random effect(s)
- ε: error term

Combining Strengths: Tree-Based Models & Random Effects

General concept: replace the fixed effect part of a mixed model by a tree or RF while keeping the modeling of the dependence structure with random effects.

Emerging Landscape:

- Promising but limited (recent) publications and implementations (*LongituRF, REEMtree, glmerTREE, MixRF, GPBoost,…*).
- *GPBoost* (based on *LightGBM*, first release in 2020): the most advanced, but rudimentary interface and limited integration with popular packages (*tidymodels, caret*).
- A Bayesian approaches (*BiMM*) show promise, but implementation released only last week...

Potential ML Algorithms

$$DILI_i \sim f(Genes + Dose + (Cluster)) + \gamma\, Drug_i$$

Elastic Net (*glmnet*): powerful tool for regression analysis, particularly when dealing with high-dimensional data and to avoid overfitting.

- Reduce Overfitting & Identify Relevant Genes: Combine L1 (Lasso) & L2 (Ridge) → Prevents the model to overfit to the training data.
- Interactions have to be specified (formula)
- ~~Random effect~~

Tree-based methods:

- **Benefit of tree-based methods:** don't need to explicitly define interactions between genes, as these methods can automatically capture them.
- **Random forest (*ranger*):** Combines multiple independent decision trees. ~~Random effect.~~
- **Gradient boosting machine (*lightGBM*):** Employs a sequential ensemble of decision trees to improve accuracy. ~~Random effect.~~
- **Gradient boosting machine and mixed effects models (*GPBoost*):** *lightGBM + random effect*

DILI Prediction: Limited Improvement with Clusters & Overfitting

Cluster	Method	.metric	Test	Training
No	Elastic Net	bal_accuracy	0.49	0.71
No	Random Forest	bal_accuracy	0.52	
No	LightGBM	bal_accuracy	0.52	0.82
No	GPBoost	bal_accuracy	0.50	1.00
No	Elastic Net	roc_auc	0.49	0.80
No	**Random Forest**	**roc_auc**	**0.56**	
No	LightGBM	roc_auc	0.53	0.91
No	GPBoost	roc_auc	0.52	1.00
Yes	Elastic Net	bal_accuracy	0.51	0.73
Yes	Random Forest	bal_accuracy	0.53	
Yes	LightGBM	bal_accuracy	0.52	0.96
Yes	GPBoost	bal_accuracy	0.50	1.00
Yes	Elastic Net	roc_auc	0.51	0.84
Yes	Random Forest	roc_auc	0.49	
Yes	LightGBM	roc_auc	0.47	0.99
Yes	**GPBoost**	**roc_auc**	**0.56**	**1.00**

- The predictive power of tested models/algorithms is limited
- Including cluster information (A/B/C) did not significantly improve the model's ability to predict DILI.
- GPBoost achieved the highest ROC AUC (with RF)
- Perfect predictions for Gpboost on the training data: drug-to-drug random effect explains the training data very well (no random effect estimated for the test data) => is there a common toxicity signal in the data ?

Pathway Analysis: Feature transformation leveraging biological knowledge

Limited Model Performance: Signal or Overfitting?

- Lack of a strong signal linking genes to DILI.
- Overfitting on irrelevant gene associations.

Overfitting Concerns: Current method using DE on 18,574 genes might capture spurious correlations.

Addressing Overfitting with Pathways:

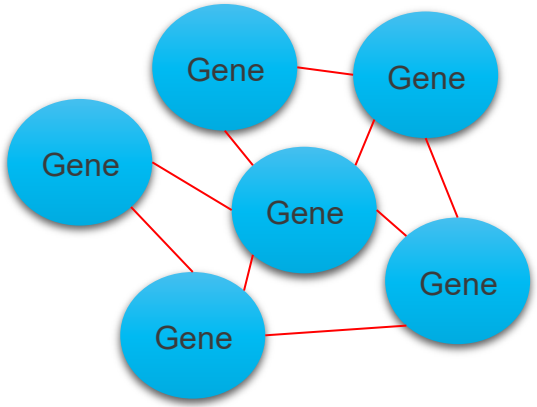
Genes usually don't function independently, they operate in networks (pathways).

Modeling these pathways, instead of individual genes:

- Avoid overfitting on noise
- Add biological relevance!

IPA, MsigDB, AOP,… databases provide curated gene-pathway links.

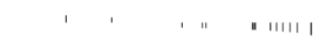
Introducing: Feature transformation using Gene Set Enrichment Analysis



ABBUD_LIF_SIGNALING_1_DN	'AHNAK' 'ALCAM' 'ANKRD40' 'ARID1A' 'BCKDHB' 'C16orf89' ...
ABBUD_LIF_SIGNALING_1_UP	'ACAA2' 'ALDOC' 'ANXA8L1' 'BCL3' 'CEBPB' 'CXCL14' ...
ABBUD_LIF_SIGNALING_2_DN	'CGA' 'CITED2' 'NALCN' 'PITX2' 'PTHLH' 'SCN1A' ...
ABBUD_LIF_SIGNALING_2_UP	'ATP1B1' 'COL11A1' 'DAB2' 'DCN' 'DIO2' 'EZR' ...
ABDELMOHSEN_ELAVL4_TARGETS	'BCL2' 'CAB39' 'CASP3' 'CDC42' 'CDH2' 'DLG4' ...
ABDULRAHMAN_KIDNEY_CANCER_VHL_DN	'ACTA2' 'ALDH1A1' 'ALDH3B1' 'ITGB3BP' 'MPPE1' 'MTMR3' ...
ABE_INNER_EAR	'APOD' 'ATP1A2' 'C19orf53' 'CA14' 'CCS' 'CLU' ...

Gene Set Enrichment Analysis (GSEA)

GSEA analyzes groups of genes to identify biologically relevant processes or pathways associated with different conditions.

Pathway	Gene ranks	NES	pval	padj
5990980_Cell_Cycle		2.66	$2.5 \cdot 10^{-26}$	$1.5 \cdot 10^{-23}$
5990979 Cell Cycle, Mitotic		2.74	$9.4 \cdot 10^{-26}$	$2.7 \cdot 10^{-23}$
5991421_Effects_of_PIP2_hydrolysis		-1.97	$2.6 \cdot 10^{-3}$	$1.6 \cdot 10^{-2}$
5991751_TCR_signaling		-1.78	$2.5 \cdot 10^{-3}$	$1.6 \cdot 10^{-2}$

Key Steps:

1.Rank the Genes:

1. Identify genes differentially expressed between conditions (drug treatment vs. control).
2. Rank these genes based on their expression changes (alternatively could be done on ranked signed p-value).

2.Enrichment Score:

1. GSEA calculates a score (NES) for each pathway.
2. NES reflects how enriched the pathway is at the top/bottom of the ranked gene list.

3.Interpretation:

1. Positive NES suggests the pathway is up-regulated.
2. Negative NES suggests the pathway is down-regulated.

Gene-Clustering using GSEA: slight improvement

Second approach: we reduced our 18,574 genes features to 68 pathways enrichment scores features using Gene Set Enrichment Analysis (GSEA) based on the AOP (Adverse Outcome Pathway) pathways.

Cluster	Method	.metric	Test	Training
No	Elastic Net	bal_accuracy	0.50	0.64
No	Random Forest	bal_accuracy	0.52	0.99
No	LightGBM	bal_accuracy	0.51	0.85
No	GPBoost	bal_accuracy	0.50	1.00
No	Elastic Net	roc_auc	0.49	0.72
No	Random Forest	roc_auc	0.52	1.00
No	LightGBM	roc_auc	0.52	0.93
No	GPBoost	roc_auc	0.53	1.00
Yes	Elastic Net	bal_accuracy	0.52	0.66
Yes	Random Forest	bal_accuracy	0.56	0.99
Yes	LightGBM	bal_accuracy	0.54	0.88
Yes	GPBoost	bal_accuracy	0.50	1.00
Yes	Elastic Net	roc_auc	0.50	0.74
Yes	Random Forest	roc_auc	0.55	1.00
Yes	**LightGBM**	**roc_auc**	**0.59**	**0.96**
Yes	GPBoost	roc_auc	0.53	1.00

Similar conclusion as feature reduction using differential expression BUT highest AUC score observed so far!

Conclusion and Discussion

Model Predictive Performance:

- Clustering information did not significantly improve the models
- Our models achieved comparable (low) performance to previous studies (Chierici et al., 2020)

Limitations of the Current Approach:

- **Limited signal in the data:** Transcriptomic data might not explain DILI toxicity alone
- **Patient Specificity:** Using data from a single donor population limits generalizability. DILI events can be rare.
- **DILI Annotation Variability:** Inconsistencies in DILI labelling across datasets (drug labels vs. literature) can introduce noise

Future Directions in the general use of ML:

- **Random Effects:** Promising potential, still require further development for user-friendly tools and broader adoption
- **Exploring Pathway Enrichment:** Leveraging pathway knowledge for feature "transformation" is promising. Further exploration in this direction is needed.

Contact and Acknowledgment

Thomas de Marchin
Associate Director, Statistics & Data Science
Thomas.demarchin@pharmalex.com

https://www.linkedin.com/in/tdemarchin/

cencora PharmaLex

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Lira Pi

Milana Filatenkova

Shreyas Hirway

Fetene Tekle

Xian Yao

Key References

1. Minerali, Eni, Daniel H. Foil, Kimberley M. Zorn, Thomas R. Lane, and Sean Ekins. “Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI).” *Molecular Pharmaceutics* 17, no. 7 (July 6, 2020): 2628–37. https://doi.org/10.1021/acs.molpharmaceut.0c00326.

2. Guo, Wenjing, Jie Liu, Fan Dong, Meng Song, Zoe Li, Md Kamrul Hasan Khan, Tucker A Patterson, and Huixiao Hong. “Review of Machine Learning and Deep Learning Models for Toxicity Prediction.” Experimental Biology and Medicine, December 6, 2023, 15353702231209421. https://doi.org/10.1177/15353702231209421.

3. Chierici, Marco, Margherita Francescatto, Nicole Bussola, Giuseppe Jurman, and Cesare Furlanello. “Predictability of Drug-Induced Liver Injury by Machine Learning.” Biology Direct 15, no. 1 (December 2020): 3. https://doi.org/10.1186/s13062-020-0259-4.

4. Hu, Jianchang, and Silke Szymczak. “A Review on Longitudinal Data Analysis with Random Forest.” Briefings in Bioinformatics 24, no. 2 (March 19, 2023): bbad002. https://doi.org/10.1093/bib/bbad002.

5. Ngufor, Che, Holly Van Houten, Brian S. Caffo, Nilay D. Shah, and Rozalina G. McCoy. “Mixed Effect Machine Learning: A Framework for Predicting Longitudinal Change in Hemoglobin A1c.” Journal of Biomedical Informatics 89 (January 2019): 56–67. https://doi.org/10.1016/j.jbi.2018.09.001.

6. https://github.com/fabsig/GPBoost