Supplementary Information

ChemBioHepatox: Multimodal Integrating Chemical Structure and Biological Fingerprint for Robust and Interpretable Hepatotoxicity Prediction

Yingqing Shou[1,2], Yongbao Fang[2,4], Yu Ma[2], Yuan Chi[2], Jingzhi Yao[2], Jie Xu[5], YongFeng Deng*[1], Bo Peng*[3], Mingliang Fang*[2]

*Corresponding authors. Email: yfdeng@seu.edu.cn, pengbo@pku.edu.cn and mlfang@fudan.edu.cn

[1] Key Laboratory of Environmental Medicine Engineering, Ministry of Education, School of Public Health, Southeast University, Nanjing 210009, China; [2] Department of Environmental Science and Engineering, Fudan University, Shanghai 200438, China; [3] SKL-ESPC and College of Environmental Sciences and Engineering, Center for Environment and Health, Peking University, Beijing 100871, China; [4] Department of Gynecology and Obstetrics, The Second Affiliated Hospital of Anhui Medical University, Hefei, 230601, China; [5] Fudan University, Shanghai 200438, China.

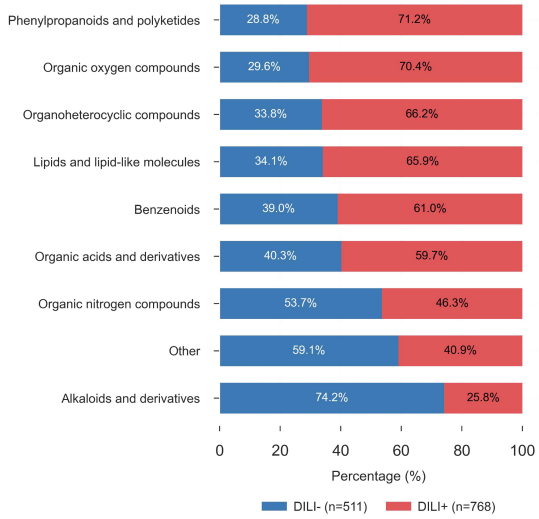
Table of Contents:

Supplementary Methods

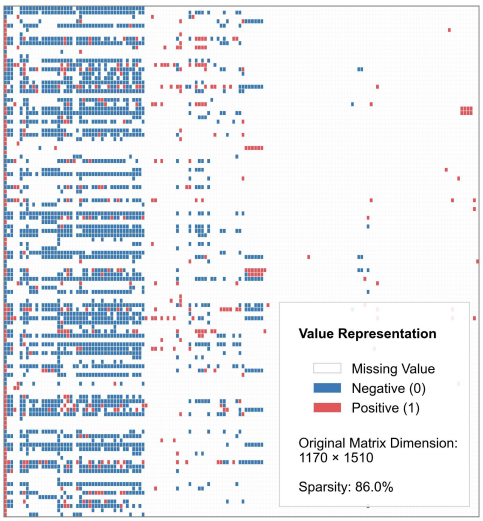
Identification of a comprehensive hepatotoxicity assay response spectrum - Step 3: Mechanism Mapping. To establish a comprehensive and systematic hepatotoxicity assay response spectrum, we further processed the 38 significantly correlated assays by eliminating redundant assays and phenotypic indicators while prioritizing molecular mechanism level assays. This refinement process involved eliminating duplicate significance assays (retaining ER-beta Agon qHTS: Summary while removing ER-beta Agon qHTS), excluding phenotypic indicators (SGOT, SGPT) in favor of mechanism-level assays, and removing weakly correlated or hepatotoxicity-irrelevant assays (such as D3 Receptor Agon qHTS). This process yielded 17 key assays. We next mapped these assays to the established comprehensive framework of hepatotoxicity mechanisms, which systematically categorizes hepatotoxicity into 12 fundamental mechanisms. Upon mapping our assays to this framework, we identified coverage gaps in two critical mechanisms: transport function disruption and cytoskeletal disruption. To achieve complete mechanistic coverage, we strategically incorporated two additional assays—MDR-1 (transmembrane transport) and ER stress (cytoskeletal function)—resulting in a final panel of 19 key assays that collectively span all 12 fundamental hepatotoxicity mechanisms.

Supplementary Figures

a.



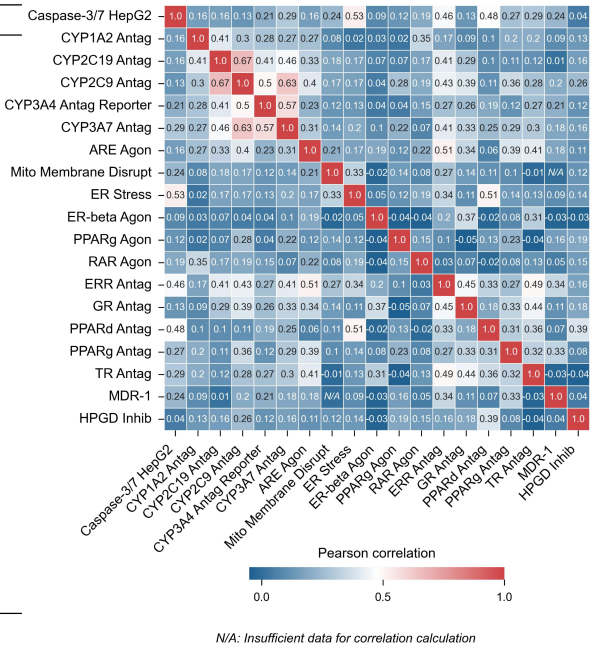
b.



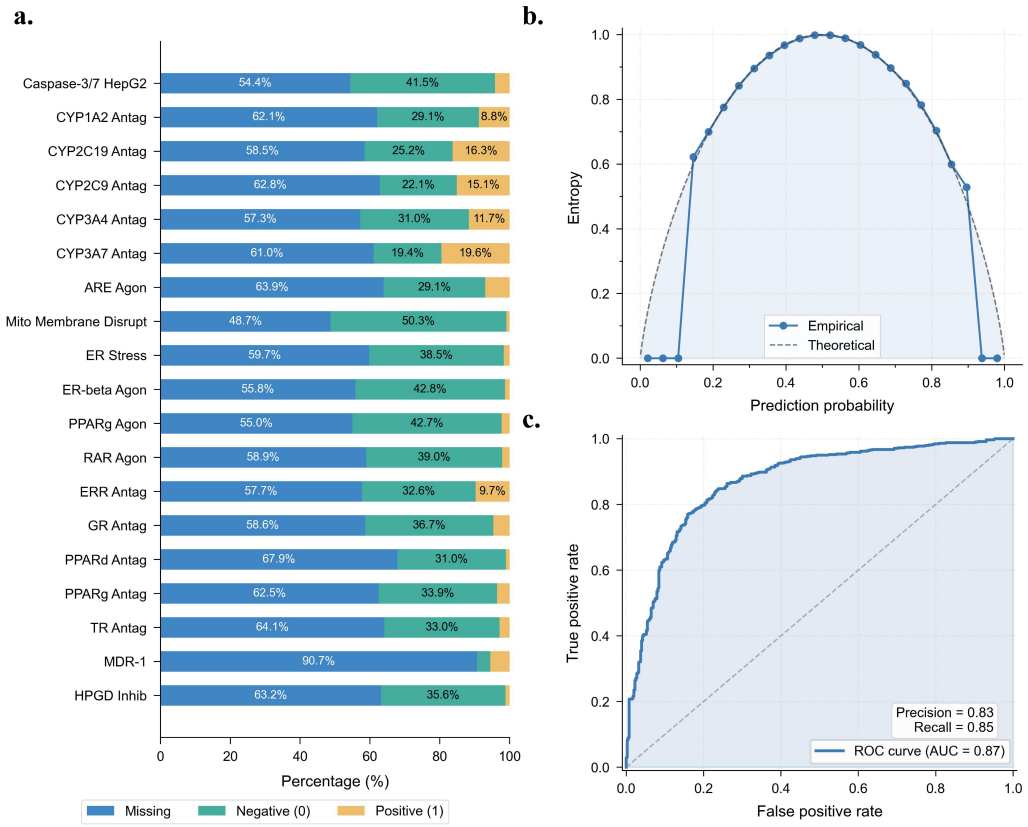
c.

Intersection：38assay	
CYP2C9 Antag qHTS	Luciferase Inhib qHTS
CYP2C9 Antag Cell qHTS	CYP3A4 Antag Reporter qHTS
CYP3A7 Antag Cell qHTS	Caspase-3/7 HepG2 qHTS
Toxo IFN-y HTS	ER-beta Agon qHTS: Summary
PPARg Agon qHTS: Summary	ER-beta Agon qHTS
ERR Antag qHTS	AR Antag MDA qHTS
D3 Receptor Agon qHTS	TR Antag qHTS
D3 Receptor Potent qHTS	ARE Agon qHTS
FDA HLAED SGOT	TR Antag Summary qHTS
Shh Antag qHTS	RAR Agon qHTS
CYP3A4 Antag Cell qHTS	PPARg Antag Summary qHTS
HPGD Inhib qHTS	PPARd Antag qHTS
FDA HLAED SGPT	FDA HLAED Liver Enzyme
LtaS Inhib HTS	PR Antag qHTS
CYP1A2 Antag qHTS	CYP2C19 Antag qHTS
Caspase-3/7 CHO qHTS	GR Antag qHTS
AR Antag qHTS	ERR Antag Cell Viab qHTS
MMP	NSD2 Inhib qHTS
PPARg Agon qHTS	RECQ1 Inhib qHTS

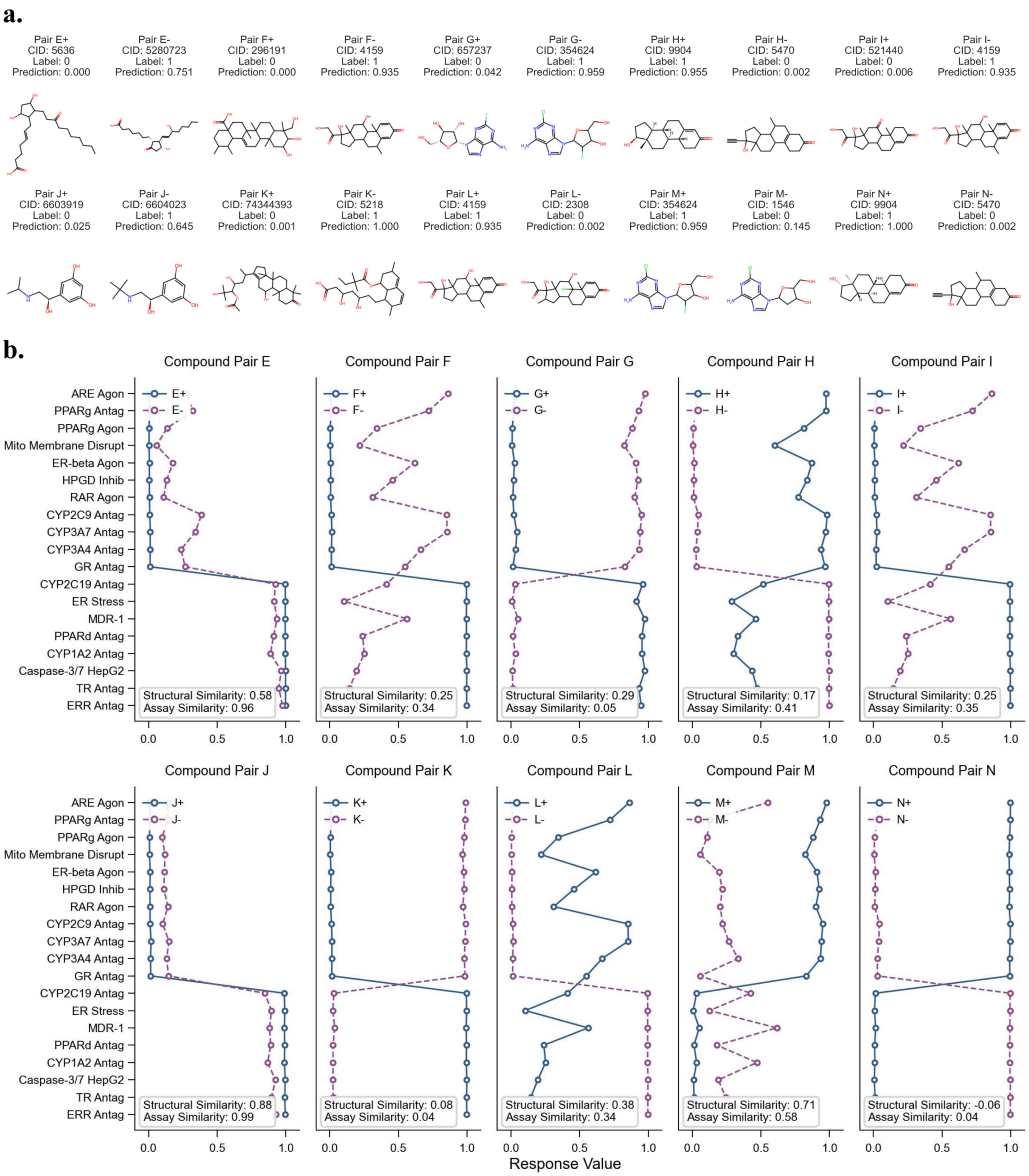
d.



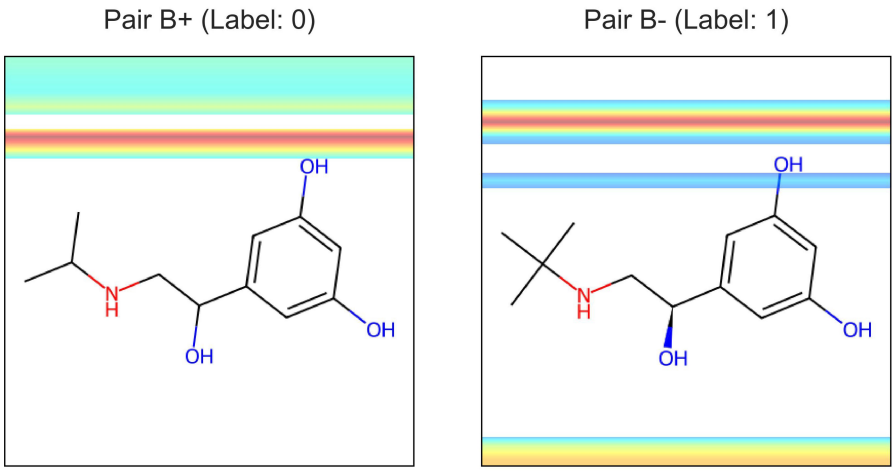
Supplementary Fig. 1 | Chemical classification and assay response analysis of hepatotoxic compounds. a Distribution of DILI-positive (red, n=768) and DILI-negative (blue, n=511) compounds across different chemical classes according to ClassyFire taxonomic system. Percentages indicate proportion within each class. **b** Visualization of the drug-assay response matrix (1170 × 1510) showing distribution of positive (red), negative (blue), and missing values (white). Matrix exhibits 86% sparsity. **c** List of 38 statistically significant hepatotoxicity-associated assays identified through dual statistical screening approach. **d** Heatmap of Pearson correlation coefficients between the 19 key assays in the final panel. Color intensity indicates correlation strength from 0.0 (blue) to 1.0 (red), with gray cells representing insufficient data for correlation calculation.



Supplementary Fig. 2 | Data characteristics and model performance metrics. a Percentage distribution of missing values and class labels across the 19 key assays. Blue segments represent missing data, green segments represent negative class (0) samples, and orange segments represent positive class (1) samples. **b** Comparison between empirical (solid blue line with circles) and theoretical (dashed gray line) entropy distributions across prediction probability values. **c** Receiver operating characteristic (ROC) curve for the standardized SMILES representation approach with area under curve (AUC), precision, and recall values indicated in the legend.

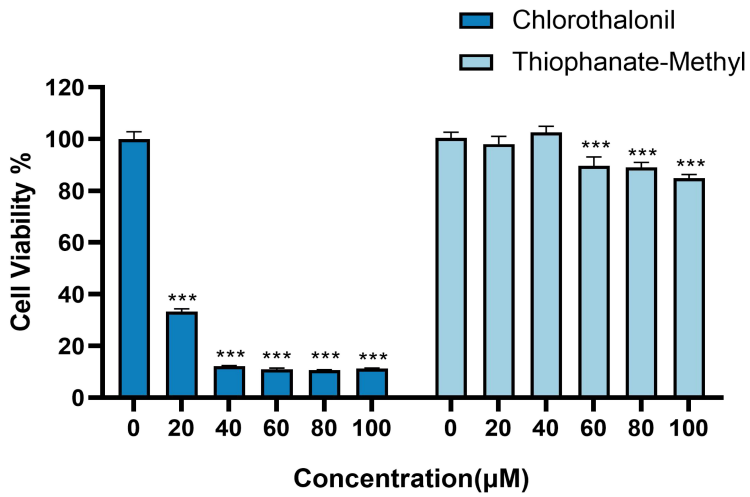


Supplementary Fig. 3 | Extended activity cliff analysis demonstrating structure-mechanism complementarity. a Molecular structures of the remaining 10 activity cliff pairs (Pairs E - N) not shown in the main text, with CID numbers, toxicity labels, and prediction probabilities. **b** Assay response profiles for 10 additional activity cliff pairs (Pairs E-N), showing response values (0.0-1.0) across 19 key assays. Blue lines with circular markers represent non-hepatotoxic compounds, while pink lines with circular markers represent hepatotoxic compounds. Structural similarity and assay similarity scores are provided at the bottom of each panel.

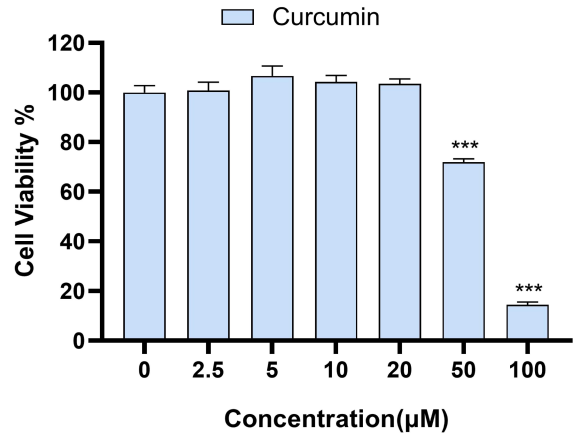


Supplementary Fig. 4 | ChemBioHepatox attention mechanism visualization for Pair B highlighting the differential attention distribution patterns between structurally similar molecules with opposite toxicity labels (B+: non-hepatotoxic, Label 0; B-: hepatotoxic, Label 1). Color intensity represents attention weight, with warmer colors indicating regions of higher attention by the model.

a.



b.



Supplementary Fig. 5 | Cell viability assays for model-predicted hepatotoxic compounds. **a** Bar graph showing dose-dependent cytotoxicity of chlorothalonil (dark blue) and thiophanate-methyl (light blue) in HepG2 cells after 24-hour exposure. Cell viability is expressed as percentage of control (0 μM), with significant decreases observed at all tested concentrations for chlorothalonil and at concentrations ≥ 60 μM for thiophanate-methyl (***$p < 0.001$). **b** Bar graph showing the dose-dependent effects of curcumin on HepG2 cell viability after 24-hour exposure. Curcumin maintains cell viability at lower concentrations (≤ 20 μM) but exhibits significant cytotoxicity at higher concentrations (50 μM and 100 μM) (***$p < 0.001$).

Supplementary Table

Supplementary Table 1. Characteristics of the 19 key assays in the hepatotoxicity response spectrum

Assay Name	Abbreviation	PubChem AID	Protein Target	Corresponding KC*
Caspase-3/7 induction in HepG2 cells by small molecules, qHTS assay: Summary	Caspase-3/7 HepG2	1347034	caspase-3; caspase 7; apoptosis-related cysteine protease	KC2; KC3
Cytochrome P450 Family 1 Subfamily A Member 2 (CYP1A2) small molecule antagonists: luciferase reporter qHTS assay	CYP1A2 Antag	1671199	cytochrome P450; family 1; subfamily A; polypeptide 2	KC1; KC12
Cytochrome P450 Family 2 Subfamily C Member 19 (CYP2C19) small molecule antagonists: luciferase reporter qHTS assay	CYP2C19 Antag	1671197	cytochrome P450; family 2; subfamily C; polypeptide 19	KC1; KC12
Cytochrome P450 Family 2 Subfamily C Member 9 (CYP2C9) small molecule antagonists: luciferase reporter qHTS assay	CYP2C9 Antag	1671198	cytochrome P450 family 2 subfamily C polypeptide 9	KC1; KC12
Cytochrome P450 Family 3 Subfamily A Member 4 (CYP3A4) small molecule antagonists: luciferase reporter qHTS assay	CYP3A4 Antag	1671201	cytochrome P450 family 3 subfamily A polypeptide 4	KC1; KC12
Cytochrome P450 family 3 subfamily A member 7 (CYP3A7) small molecule antagonists: luciferase cell-based qHTS	CYP3A7 Antag	1963596	cytochrome P450 3A7 [Homo sapiens]	KC1; KC12

assay				
qHTS assay for small molecule agonists of the antioxidant response element (ARE) signaling pathway: Summary	ARE Agon	743219	nuclear factor erythroid 2-related factor 2 isoform 1	KC5; KC8
qHTS assay for small molecule disruptors of the mitochondrial membrane potential - cell viability	Mito Membrane Disrupt	720634	NA	KC2; KC5; KC7
qHTS assay to identify small molecule agonists of the endoplasmic reticulum stress response signaling pathway - cell viability counter screen	ER Stress	1159517	NA	KC2; KC8; KC10
qHTS assay to identify small molecule agonists of the estrogen receptor beta (ER-beta) signaling pathway: Summary	ER-beta Agon	1259394	estrogen receptor 2 (ER beta)	KC3; KC12
qHTS assay to identify small molecule agonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway: Summary	PPARg Agon	743140	peroxisome proliferator activated receptor gamma	KC3; KC12
qHTS assay to identify small molecule agonists of the retinoic acid receptor (RAR) signaling pathway	RAR Agon	1159553	retinoic acid nuclear receptor alpha variant 1	KC3; KC9
qHTS assay to identify small molecule antagonists of the estrogen related receptor (ERR) signaling pathway from Tox21 library	ERR Antag	1224848	estrogen-related nuclear receptor alpha	KC1; KC12

qHTS assay to identify small molecule antagonists of the glucocorticoid receptor (GR) signaling pathway: Summary	GR Antag	720725	glucocorticoid receptor [Homo sapiens]	KC6; KC12
qHTS assay to identify small molecule antagonists of the peroxisome proliferator-activated receptor delta (PPARd) signaling pathway	PPARd Antag	743215	peroxisome proliferator-activated receptor delta	KC3; KC12
qHTS assay to identify small molecule antagonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway: Summary	PPARg Antag	743199	peroxisome proliferator activated receptor gamma	KC8; KC12
qHTS assay to identify small molecule antagonists of the thyroid receptor (TR) signaling pathway: Summary	TR Antag	743167	thyroid hormone receptor beta isoform 2	KC1; KC12
MDR-1	MDR-1	377	ATP-dependent translocase ABCB1 isoform 2	KC4; KC7
qHTS Assay for Inhibitors of HPGD (15-Hydroxyprostaglandin Dehydrogenase)	HPGD Inhib	894	15-hydroxyprostaglandin dehydrogenase [NAD(+)] isoform 1	KC6; KC11

* KC (Key Characteristics) of Hepatotoxicants:
KC1. Is reactive and/or is metabolized (bioactivated) to reactive moieties
KC2. Causes death (apoptosis and/or necrosis) of liver cells
KC3. Affects liver cell proliferation and/or tissue regeneration
KC4. Disrupts transport function
KC5. Induces oxidative stress

KC6. Triggers immune-mediated responses in liver
KC7. Causes mitochondrial dysfunction
KC8. Activates stress signaling pathways
KC9. Causes cholestasis
KC10. Disrupts cellular cytoskeleton
KC11. Causes liver fibrosis
KC12. Disrupts liver metabolism, including of lipids and proteins

Supplementary Table 2. Characteristics of the 19 key assays in the hepatotoxicity response spectrum

Pair ID	Assay Similarity	Structural Similarity	Pair ID	Assay Similarity	Structural Similarity
Pair 1	1.00	0.72	**Pair 20**	0.98	0.53
Pair 2	1.00	0.94	**Pair 21**	1.00	0.42
Pair 3	1.00	0.29	**Pair 22**	1.00	0.75
Pair 4	1.00	0.36	**Pair 23**	0.69	0.15
Pair 5	1.00	0.36	**Pair 24**	1.00	0.62
Pair 6	0.35	0.13	**Pair 25**	0.99	0.61
Pair 7	1.00	0.47	**Pair 26**	1.00	0.66
Pair 8	1.00	0.86	**Pair 27**	0.98	0.65
Pair 9	1.00	0.89	**Pair 28**	1.00	0.26
Pair 10	0.12	0.57	**Pair 29**	0.54	0.77
Pair 11	1.00	0.23	**Pair 30**	1.00	0.70
Pair 12	1.00	0.64	**Pair 31**	0.67	0.15
Pair 13	0.68	0.13	**Pair 32**	0.76	0.82
Pair 14	1.00	0.93	**Pair 33**	1.00	0.80
Pair 15	1.00	0.48	**Pair 34**	0.10	0.59
Pair 16	1.00	0.54	**Pair 35**	0.91	0.76
Pair 17	1.00	0.51	**Pair 36**	1.00	0.88
Pair 18	1.00	0.64	**Pair 37**	0.85	0.56
Pair 19	1.00	0.55	**Pair 38**	0.07	0.27

Supplementary Table 3. Predicted hepatotoxicity of synthetic food colorants and supporting evidence

Food Colorant	E Number	Predicted Probability	Key Findings from Literature	Reference
Sunset Yellow	E110	0.99	Hepatocyte damage; elevated liver enzymes; altered antioxidant systems	27
Carmoisine	E122	1.00	Oxidative stress; elevated liver enzymes; histopathological changes	28
Allura Red	E129	0.99	Oxidative stress; inflammatory responses; dose-dependent effects	29
Tartrazine	E102	1.00	Increased oxidative markers; altered liver function	30
Curcumin	E100	0.95	Dose dual effect: hepatoprotective at low doses, hepatotoxic at high doses	31
Cochineal Red A	E124	0.96	Mitochondrial dysfunction; increased oxidative stress	32
Brilliant Blue	E133	1.00	Liver tissue accumulation; vacuolation, swelling, necrosis and pyknosis of liver cells	33