Supplementary Information (SI) for Green Chemistry.
This journal is © The Royal Society of Chemistry 2025

Supporting Information-Annex 1

Toward Comprehensive Scientific Information on Plastic-Related Chemicals Powered by Artificial Intelligence

Kai Zhao[$, a], Xiting Peng[$, a], Ran Tao[$, a], Yuchen Su [c], ShiYue Huang [c], Hang Fu [d], Xiaonan Wang*[, a, b], Shanying Hu*[, a]

[a]*Department of Chemical Engineering, Tsinghua University, Beijing 100084, P. R. China*

[b] *Institute for Carbon Neutrality, Tsinghua University, Beijing 100084, P. R. China*

[c] *Department of Computer Science, University of Electronic Science and Technology of China, Chengdu, Sichuan 610056, P. R. China*

[d] *School of Resources and Environment, Nanchang University, Nanchang, 330031, P. R. China*

[$] *Kai Zhao, Xiting Peng, and Ran Tao contributed equally to this paper*

Corresponding Authors:

wangxiaonan@tsinghua.edu.cn (X. Wang), hxr-dce@tsinghua.edu.cn (S. Hu).

The Supporting Information consists of two parts: Annex 1 (this document), which presents detailed methods and results, and Annex 2 (Excel file), which contains the finalized plastic-related chemical database. This Annex 1 file contains 108 pages, 22 supplementary figures and 24 supplementary tables.

Table of Contents

1. Details of functional labels of plastic additives

1.1 Second-level functional labels (major categories)

Based on the distinct functional attributes that plastic additives provide to plastics, the second-level functional labels were defined as the following 24 classes.

Table S1. Second-level functional labels of plastic additives.

Num.	Second-level functional labels	Functional descriptions
A01	Plasticizer	Enhances flexibility and workability of plastic materials by reducing intermolecular forces.
A02	Antioxidant	Prevents or slows down oxidation to protect polymers from degradation during processing and use.
A03	Light stabilizer	Shields plastics from UV-induced degradation, preserving appearance and mechanical properties.
A04	Flame retardant	Reduces flammability and delays ignition or flame propagation in plastic products.
A05	Curing agent and curing accelerators	Initiate or accelerate crosslinking reactions during polymer curing processes.
A06	Heat stabilizer	Protects polymers from thermal degradation during processing and high-temperature use.
A07	Colorants	Impart color to plastics for aesthetic or functional purposes.
A08	Coupling agent	Enhances adhesion between dissimilar materials, improving mechanical strength and durability.
A09	Crosslinking agent	Promotes the formation of chemical bonds between polymer chains to increase strength and stability.
A10	Photo Initiator	Generates reactive species upon light exposure to start polymerization or curing reactions.
A11	Antimicrobial agent	Inhibits the growth of bacteria, fungi, and other microbes on plastic surfaces.
A12	Polymerization inhibitor	Prevents unwanted or premature polymerization during processing or storage.

Num.	Second-level functional labels	Functional descriptions
A13	Chemical Foaming Agents	Produces gas during processing to create a cellular structure in foamed plastic products.
A14	Antistatic agents	Reduces surface static charge buildup to prevent dust attraction and electrical discharge.
A15	Fluorescent whitener	Absorbs UV light and emits blue light to enhance brightness and reduce yellowing.
A16	Drip and mist eliminators	Minimizes dripping or misting by altering surface tension or flow characteristics.
A17	Lubricants	Reduces friction between polymer surfaces during processing to improve flow and release.
A18	Nucleator	Promotes the formation of crystalline structures during polymer solidification to enhance mechanical properties.
A19	Impact Modifiers	A type of chemical that improves the low-temperature embrittlement of polymer materials and gives them greater toughness.
A20	Anti-sticking agent	Prevents plastic surfaces from adhering to equipment or each other during processing.
A21	Releasing agent	A type of chemical used to prevent other materials from bonding to surfaces
A22	Acid binding agent	Neutralizes acidic degradation products to stabilize polymer performance.
A23	Slipping agent	A type of chemical that reduces the coefficient of friction on plastic surfaces to the required level or value.
A24	Others	/

1.2 Third-level functional labels (subcategories)

Based on differences in characteristic functional groups, detailed descriptions of the 83 third-level functional labels are provided, along with the number of manually annotated data entries collected for each.

Table S2. Third-level functional labels of plastic additives.

Num.	Third-level functional labels	Count of manually annotated data
C00	Plasticizer_Phthalate	33
C01	Plasticizer_Terephthalate	1
C02	Plasticizer_Isophthalate	1
C03	Plasticizer_Adipic acid esters	12
C04	Plasticizer_Azelaic acid esters	2
C05	Plasticizer_Fumaric acid esters	2
C06	Plasticizer_Citric acid esters	5
C07	Plasticizer_Trimellitate	4
C08	Plasticizer_Itaconic acid esters	2
C09	Plasticizer_Lauric acid esters	2
C10	Plasticizer_Maleic acid esters	4
C11	Plasticizer_Oleate	5
C12	Plasticizer_Sebacic acid esters	7
C13	Plasticizer_Stearate	3
C14	Plasticizer_Sulfonic acid derivatives	3

Num.	Third-level functional labels	Count of manually annotated data
C15	Plasticizer_Glycol derivatives, glycerol derivatives, propylene glycol derivatives and other polyol derivatives	17
C16	Plasticizer_Epoxy derivatives	5
C17	Plasticizer_Phosphoric acid	12
C18	Plasticizer_Chlorine Plasticizer	1
C19	Plasticizer_Polymeric plasticizers	4
C20	Plasticizer_Other types	24
C21	Antioxidant_Amines	15
C22	Antioxidant_Bisphenol monoacrylate	2
C23	Antioxidant_Hindered phenolics	54
C24	Antioxidant_Metal passivator	4
C25	Antioxidant_Phosphite	14
C26	Antioxidant_Thioether	10
C27	Antioxidant_Triazine	6
C28	Antioxidant_Other types	2
C29	Light stabilizer_Benzoate	4
C30	Light stabilizer_Benzophenones	13
C31	Light stabilizer_Benzpropyltriazole	14

Num.	Third-level functional labels	Count of manually annotated data
C32	Light stabilizer_Cyanoacrylates	2
C33	Light stabilizer_Hindered amines	20
C34	Light stabilizer_Hydroxybenzotriazines	3
C35	Light stabilizer_Light shielding agents	2
C36	Light stabilizer_Nickel-containing compounds	4
C37	Light stabilizer_Salicylate esters	3
C38	Light stabilizer_Other types	8
C39	Flame retardant_Halogenated Flame retardant	43
C40	Flame retardant_Inorganic flame retardant	11
C41	Flame retardant_Phosphorus Flame retardant	21
C42	Flame retardant_Other types	4
C43	Curing agents and curing accelerators_Acid anhydride	10
C44	Curing agents and curing accelerators_Amines	33
C45	Curing agents and curing accelerators_Other Types	9
C46	Heat stabilizer_Inorganic and organic lead salts	5
C47	Heat stabilizer_Lead diformate	3
C48	Heat stabilizer_Metal soaps and metal salts	16
C49	Heat stabilizer_Organic primary and secondary stabilizers	10

Num.	Third-level functional labels	Count of manually annotated data
C50	Heat stabilizer_Organic tin	8
C51	Colorants_Inorganic Colorants	14
C52	Colorants_Organic colorants	28
C53	Coupling agent_Organic Chromium	1
C54	Coupling agent_Silanes	36
C55	Crosslinking agent_Organic peroxides	22
C56	Crosslinking agent_Other types	12
C57	Photo Initiator_Photoinitiator	22
C58	Photo Initiator_Photosensitizing aids	2
C59	Antimicrobial agent	17
C60	Polymerization inhibitor	15
C61	Chemical Foaming Agents_Azo	4
C62	Chemical Foaming Agents_Nitroso compounds	2
C63	Chemical Foaming Agents_Sulfonylhydrazine	3
C64	Chemical Foaming Agents_Other types	5
C65	Antistatic agents_Amphoteric ion type	3
C66	Antistatic agents_Anionic	1
C67	Antistatic agents_Cationic	4

Num.	Third-level functional labels	Count of manually annotated data
C68	Antistatic agents_Flammable	1
C69	Antistatic agents_Nonionic	3
C70	Antistatic agents_Polymer type	1
C71	Fluorescent whitener	12
C72	Drip and mist eliminators_Compounded Flow Drops	2
C73	Drip and mist eliminators_Fluidized droplet monomers	9
C74	Lubricants_Fatty acids and derivatives	10
C75	Nucleator	8
C76	Impact Modifiers	5
C77	Anti-sticking agent	3
C78	Releasing agent_Other types	1
C79	Releasing agent_amides	2
C80	Acid binding agent_nan	2
C81	Slipping agent_Fatty acids and derivatives	1
C82	Others	1

2. Details of LLM-based workflow for parsing chemical composition

2.1 Details of chemical-related properties retrieved from PubChem

For each chemical entry, 17 types of properties were retrieved from PubChem. The detailed records are included in the plastic-related chemical database presented in Annex 3.

Table S3. List of chemical-related properties.

Properties	Descriptions
Molecular_weight	The molecular weight is the sum of all atomic weights of the constituent atoms in a compound, measured in g/mol. In the absence of explicit isotope labelling, averaged natural abundance is assumed. If an atom bears an explicit isotope label, 100% isotopic purity is assumed at this location.
Melting point	Melting point (freezing point) (M.P.) is the temperature at which a crystal exists in a solid-liquid coexistence state during the process of changing its physical state from solid to liquid under atmospheric pressure.
TGA	TGA data represent the cumulative weight loss percentages of a substance at specified temperatures, indicating its thermal decomposition behavior.
XLogP	Computationally generated octanol-water partition coefficient or distribution coefficient. XLogP is used as a measure of hydrophilicity or hydrophobicity of a molecule.
ExactMass	The mass of the most likely isotopic composition for a single molecule, corresponding to the most intense ion/molecule peak in a mass spectrum.
MonoisotopicMass	The mass of a molecule, calculated using the mass of the most abundant isotope of each element.
TPSA	Topological polar surface area, computed by the algorithm described in the paper by Ertl et al [1].

Properties	Descriptions
Complexity	The molecular complexity rating of a compound, computed using the Bertz/Hendrickson/Ihlenfeldt formula.
Charge	The total (or net) charge of a molecule.
HBondDonorCount	Number of hydrogen-bond donors in the structure.
HBondAcceptorCount	Number of hydrogen-bond acceptors in the structure.
HeavyAtomCount	Number of non-hydrogen atoms.
IsotopeAtomCount	Number of atoms with enriched isotope(s)
CovalentUnitCount	Number of covalently bound units.
PatentCount	Number of patent documents linked to this compound.
PatentFamilyCount	Number of unique patent families linked to this compound (e.g. patent documents grouped by family).
LiteratureCount	Number of articles linked to this compound (by PubChem's consolidated literature analysis).

2.2 Details of prompt engineering

Considering both the model's performance and token cost, GPT-4 Turbo was selected as the preferred model. Interactions were conducted using the following prompts via API access to the GPT-4 Turbo model.

1) Question1:

Respond in JSON format with 0 or 1, where 1 means 'yes' and 0 means 'no'. Based on the text provided, answer the following three questions: Is the described substance a mixture? Does it involve unknown reaction products? Does it contain polymers?

Example of the output format:

```
{
   "mixture": 1,
   "reaction": 0,
   "polymers": 1
}
```

2) Question2:

Answer only the names in json format. What are the main ingredients in this mixture (answer with common names, Retain only the name associated with the chemical)?'

Example of the output format:

```
{
   "main_ingredients": ["xxxx"]
}
```

3) Question3:

Answer only the name. What is the common name of this substance (If you are unsure of the common name of the substance, answer "unsure".)?

2.3 Evaluation and validation of LLM- based workflow for parsing chemical composition

Each extracted result from the mixture entries was manually examined and assigned to its corresponding category in the confusion matrix (Table S4). The counts for each category were then used in Equations 1 to 4 to calculate the four performance metrics.

Table S4. Confusion Matrix of the LLM Extraction Task

	Should be extracted	Should not be extracted
LLM extracted	True Positive (TP)	False Positive (FP)
LLM not extracted	False Negative (FN)	True Negative (TN)

1) Precision:

$$Precision = \frac{TP}{TP + FP} \qquad (Eq - 1)$$

This is the proportion of correctly extracted values among all extractions by the model.

2) Recall:

$$Recall = \frac{TP}{TP + FN} \qquad (Eq - 2)$$

This is the proportion of correct values extracted by the model.

3) F1 Score:

$$F1\ Score = 2 * Recall * \frac{Precision}{Recall + Precision} \qquad (Eq - 3)$$

The F1 Score is the harmonic mean of Precision and Recall. It balances the trade-off between Precision and Recall, providing a single measure of overall model performance.

4) Accuracy

$$Accuracy = \frac{TP}{TP + FN + FP} \qquad (Eq - 4)$$

Accuracy represents the proportion of correctly extracted values (true positives) among all cases, including missed values (false negatives) and incorrect extractions (false positives).

Manual verification confirmed whether the substances extracted by the LLM matched the relevant components explicitly or implicitly mentioned in the original text.

Predictions were regarded as correct when the identified entities corresponded to reasonable reactants, monomers, or major compositional substances associated with the described system. Chemically plausible but unconfirmed cases were recorded separately as undeterminable or not found, to indicate the intrinsic difficulty of manually resolving mixture compositions.

The outcomes of this manual verification formed the basis of both Figure S1-S4 and Table S5. Figure S1-S4 visualizes the overall distribution of extraction performance, where the color intensity represents the performance and the square size indicates the number of samples within each group. Table S5 provides the corresponding quantitative statistics, listing precision, recall, F1 score, and accuracy values for each naming category and compositional complexity. Together, these results reflect how effectively the LLM achieved the intended relevance-based extraction objective, i.e., identifying substances related to the mixture or reaction system rather than reproducing the final product structures.

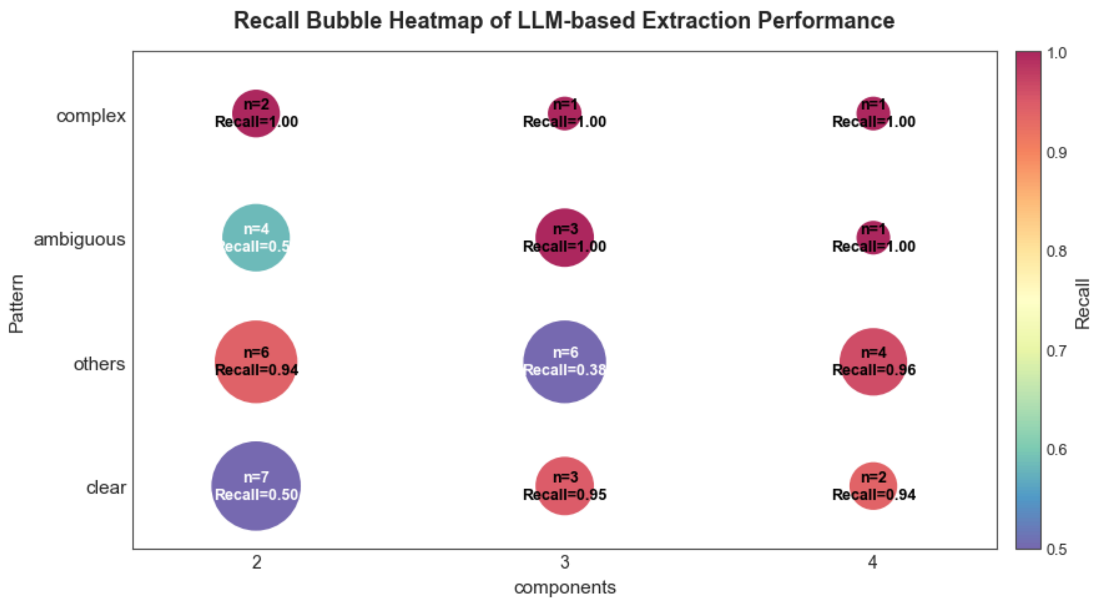

Figure S1. Heatmap of LLM-based extraction performance (recall) across naming categories and component complexities.

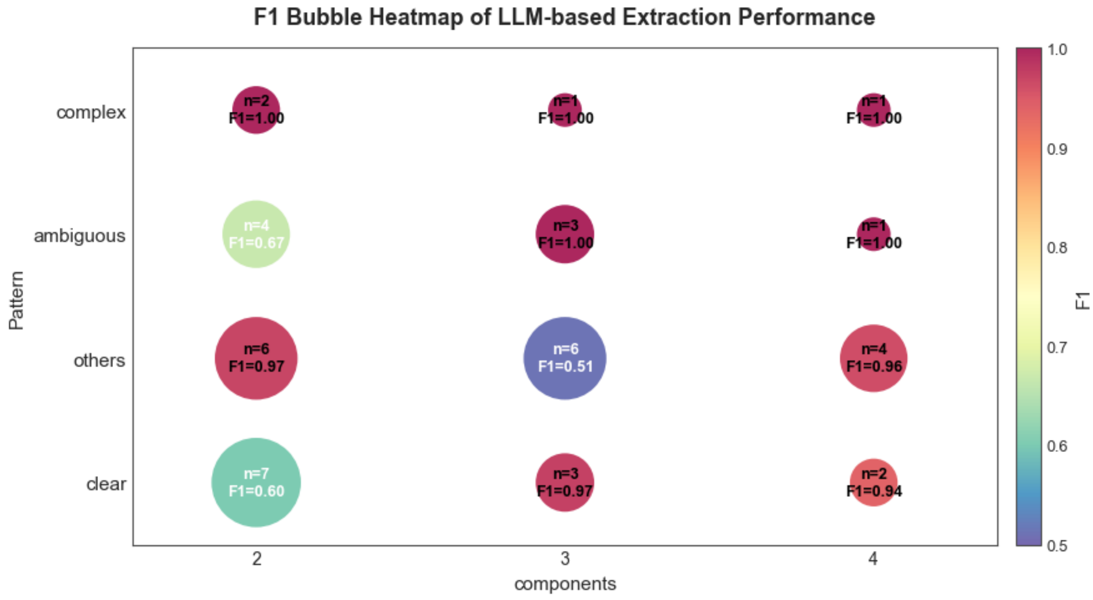

Figure S2. Heatmap of LLM-based extraction performance (F1 score) across naming categories and component complexities.

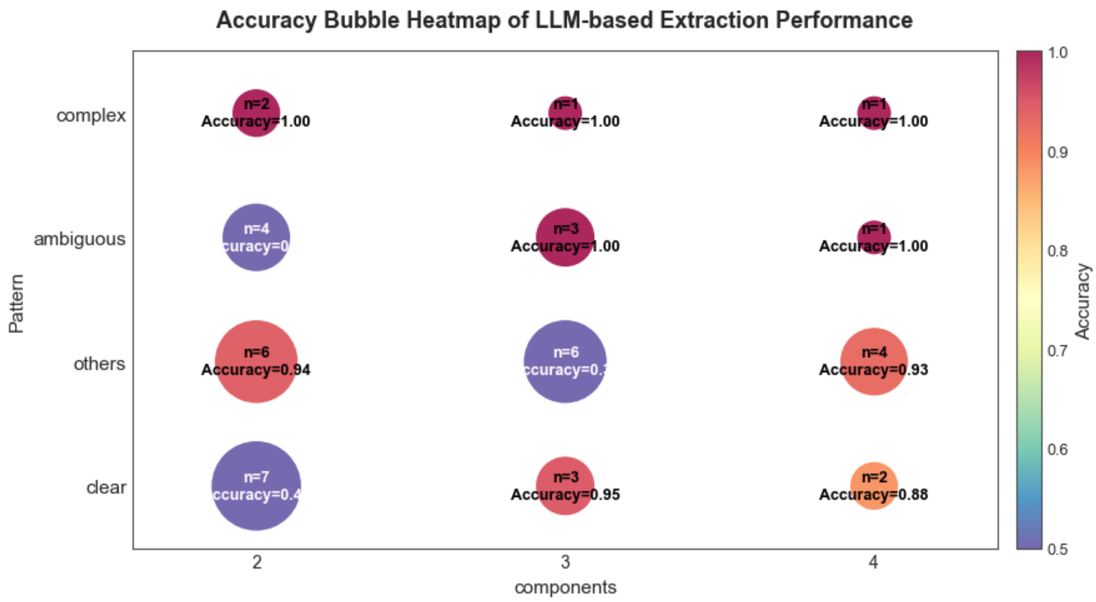

Figure S3. Heatmap of LLM-based extraction performance (Accuracy) across naming categories and component complexities.

As shown in Figure S3, the overall mean F1 value was approximately 0.80, indicating reliable extraction performance. High F1 scores (≥ 0.9) were achieved for ambiguous/regulatory and ≥4-component structural names, while the lowest performance (F1 ≈ 0.19) appeared in the 3-component complex/nested category.

Table S5. Retrieval-based validation results for LLM extraction outputs

Naming Category	Components	Precision	Recall	F1	Accuracy
Ambiguous	2	0.78	0.58	0.67	0.50
Ambiguous	3	1.00	1.00	1.00	1.00
Ambiguous	4	1.00	1.00	1.00	1.00
TOTAL		0.90	0.78	0.84	0.72
Clear	2	0.75	0.50	0.60	0.43
Clear	3	1.00	0.95	0.97	0.95
Clear	4	0.94	0.94	0.94	0.88
TOTAL		0.90	0.76	0.83	0.70
Complex	2	1.00	1.00	1.00	1.00
Complex	3	1.00	1.00	1.00	1.00
Complex	4	1.00	1.00	1.00	1.00
TOTAL		1.00	1.00	1.00	1.00
Others	2	1.00	0.94	0.97	0.94
Others	3	0.77	0.38	0.51	0.34
Others	4	0.96	0.96	0.96	0.93
TOTAL		0.93	0.74	0.82	0.70
OVERALL		0.92	0.78	0.85	0.74

Table S5 summarizes the quantitative metrics of the manual and retrieval-based validation for the 40 representative samples. The model performs excellently for clear and complex categories, achieving high F1 scores (≥0.90). It demonstrates perfect extraction in the complex category (F1 = 1.00) for all component levels, indicating robust performance for well-defined nomenclature. Challenges with Ambiguous and Multi-Component Names: Performance drops notably for 3-component complex/nested names, as shown in the others category (F1 = 0.51), highlighting difficulties in parsing more complex or ambiguous names. In conclusion, the LLM is highly effective for well-defined nomenclature but struggles with complex, multi-layered names.

2.4 Manual retrieval and validation of chemical structures

To validate the LLM-assisted extraction results, a manual retrieval process was carried out using several authoritative databases, including PubMed, SciFinder, and ECHA, which are considered reliable platforms for chemical structure retrieval. ChemicalBook and Chemical Encyclopedia were used as complementary resources. The retrieval process followed three main steps:

1) Initial Search Using CAS Numbers. When available, CAS numbers were used as the primary search method to verify whether the substance could be identified as a pure compound. This method typically allowed for the identification of pure substances associated with their CAS numbers or structures.
2) Structure Retrieval via SciFinder. For substances without a CAS number or when

additional verification was needed, SciFinder was employed to search for the corresponding chemical structures. The structures retrieved via SciFinder were cross-referenced to validate their correctness.

3) Comparison and Categorization. After retrieving the structures, they were compared to the predictions made by the LLM. The retrieval results were classified into the 5 categories.

Table S6. Comparison of results of chemically structured manual searches.

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
40	1000 85-64-1	Quaternary ammonium compounds, [2-[[2-[(2-carboxyethyl)(2-hydroxyethyl)amino]ethyl]amino]-2-oxoethyl]coco alkyldimethyl, inner salts	None	Not retrieved	Quaternary ammonium compounds
75	1003 300-73-9	Phosphoric acid, mixed esters with [1,1′-biphenyl]-4,4′-diol and phenol	ADEKA product info	Polymer mixture	Phosphoric acid Biphenyl-4,4'-diol Phenol
197	1013 56-	Barium calcium magnesium strontium	None	Not retrieved	Barium; Calcium;

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
	96-1	zinc oxide phosphate, copper-doped			Magnesium; Strontium; Zinc; Copper
1112	115019-51-7	ester of fatty acid (saturated C4-22, unsaturated C16-18) with aliphatic monohydric alcohol (saturated C2-18) and aromatic polyol ether	None	Not retrieved	fatty acid ester aliphatic; alcohol; aromatic polyol ether
2940	1571954-81-8	1,4-Benzenedicarboxylic acid, mixed Bu and 2-ethylhexyl diesters	ECHA	Three isomers (CAS 1429441-82-6, 6422-86-2, 1962-75-0)	Dibutyl phthalate; Di(2-ethylhexyl) phthalate
4053	22450-96-0	Borate(1-), bis[2-(hydroxy-╬‖O)benzoato(2-)-	Chemical Encyclop	Borate complex; tributylam	Boric acid, salicylic acid, dibutylamine

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
		╬‖O]-, (T-4)-, hydrogen, compd. with N,N-dibutyl-1-butanamine (1:1:1)	edia	ine	
5892	329211-92-9	Synthetic fibers, alumina–calcia–silica–zirconia glass	None	Not retrieved	Synthetic fibers; Aluminum oxide; Calcium oxide; Magnesium oxide; Silica
8186	610787-76-3	Hexanoic acid, 2-ethyl-, mixed triesters with benzoic acid and trimethylolpropane	None	Not retrieved	Hexanoic acid; Benzoic acid; Trimethylolpropane
8187	610787-77-4	Hexanoic acid, 2-ethyl-, mixed diesters with benzoic acid and neopentyl glycol	None	Not retrieved	Hexanoic acid; Benzoic acid; Neopentyl glycol
9485	6842	Fatty acids, C16 and C18-	ECHA	TMP	Trimethylolpro

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
	4-27-1	unsatd., triesters with trimethylolpropane		trioleate	pane trioleate
9894	68991-46-8	Linseed oil, Bu ester, epoxidized	None	Not retrieved	Linseed oil; Epoxidized butyl ester
10384	73507-36-5	2-Naphthalenesulfonic acid, 7-(benzoylamino)-4-hydroxy-3-[2-[4-[2-(4-sulfophenyl)diazenyl]phenyl]diazenyl]-, compds. with N,N'-bis(mixed Ph and tolyl and xylyl)guanidine monohydrochloride	ChemicalBook	ACI dye; guanidine; HCl	Acid Red 52, Basic Violet 14 (incorrect)
10434	742087-49-6	D-Glucopyranose, oligomeric, C10-16-alkyl glycosides, 2-hydroxy-3-sulfopropyl ethers, sodium salts	ECHA	Alkyl polyglucoside; sulfopropyl ether	Alkyl polyglucoside; Sodium hydroxy sulfopropyl ether

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
11211	8023-77-6	Resins, oleo-, capsicum	None	Not retrieved	Resins; Oleo; Capsicum
11236	8050-25-7	Resin acids and rosin acids, esters with TEG	SciFinder	Resin acids; TEG	Resin acids; Rosin acids; Triethylene glycol
11237	—	Resin acids and rosin acids, esters with pentaerythritol	ChemicalBook	Resin acids; pentaerythritol	Rosin; Pentaerythritol
11552	—	Fatty acids C16-18, esters with diethylene glycol	SciFinder	Correct mixture	Fatty acid esters
11579	—	Hexanoic acid, 2-ethyl-, mixed diesters with benzoic acid and triethylene glycol	None	Not retrieved	2-ethylhexanoic acid; benzoic acid; triethylene glycol
11632	—	mixture of methyl-	None	Not	alkanamides

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
		branched and linear C14-C18alkanamides, derived from fatty acids		retrieved	
12681	—	Phosphoric Acid, C9-11-Branched And Linear Alkyl Esters, Potassiumsalts	None	Not retrieved	Phosphoric Acid; C9-11-Branched And Linear Alkyl Esters; Potassium Salts
12682	—	Phosphoric Acid, C12-14-Branched And Linear Alkyl Esters, Potassium Salts	None	Not retrieved	Phosphoric Acid; Potassium Salts
16377	—	Formaldehyde-2-nonylphenol (1:1)	ChemNet	2-Nonylphenol; formaldehyde	Formaldehyde; nonylphenol

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
16498	—	TPAI6[MEG]5[DEG] mixture	None	Not retrieved	Tetraphenylarsonium; Monoethylene glycol; Diethylene glycol
16542	—	All salts of Al, NH4, Ba, Ca, Co, Cu, Fe, Li, Mg, Mn, K, Na, and Zn of Hexanoic acid	None	Not retrieved	Aluminum hexanoate; Ammonium hexanoate; Barium hexanoate; Calcium hexanoate; Cobalt hexanoate; Copper hexanoate; Iron hexanoate;

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
					Lithium hexanoate; Magnesium hexanoate; Manganese hexanoate; Potassium hexanoate; Sodium hexanoate; Zinc hexanoate
16792	—	POH n-C10–C35	None	Not retrieved	Petroleum Hydrocarbons
16843	—	Reaction mass of tris(2-chloropropyl) phosphate and tris(2-chloro-1-methylethyl) phosphate and Phosphoric acid, bis(2-chloro-1-methylethyl) 2-	None	Not retrieved	tris(2-chloropropyl) phosphate; tris(2-chloro-1-methylethyl) phosphate; Phosphoric

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
		chloropropyl ester and Phosphoric acid, 2-chloro-1-methylethyl bis(2-chloropropyl) ester			acid, bis(2-chloro-1-methylethyl) 2-chloropropyl ester; Phosphoric acid, 2-chloro-1-methylethyl bis(2-chloropropyl) ester
16865	—	mixture composed of 97 % tetraethyl orthosilicate (TEOS) with CAS No 78-10-4 and 3 % hexamethyldisilazane (HMDS) with CAS No 999-97-3	None	Not retrieved	tetraethyl orthosilicate; hexamethyldisilazane
17013	—	acids, C2-C24, aliphatic, linear, monocarboxylic	None	Not retrieved	fatty acids; glycerol esters

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
		from natural oils and fats, and their mono-, di- and triglycerol esters (branched fatty acids at naturally occuring levels are included)			
17047	—	All salts of Al, NH4, Ba, Ca, Co, Cu, Fe, Li, Mg, Mn, K, Na, and Zn of Propionic acid	None	Not retrieved	Aluminum propionate; Ammonium propionate; Barium propionate; Calcium propionate; Cobalt propionate; Copper propionate; Iron propionate;

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
					Lithium propionate; Magnesium propionate; Manganese propionate; Potassium propionate; Sodium propionate; Zinc propionate
17176	—	trialkyl acetic acid (C7-C17), vinyl esters	None	Not retrieved	trialkyl acetic acid; vinyl esters
17196	—	Reaction mass of Bis(1,2,2,6,6-pentamethyl-4-piperidyl) sebacate and Methyl 1,2,2,6,6-pentamethyl-4-	None	Not retrieved	Bis(1,2,2,6,6-pentamethyl-4-piperidyl) sebacate; Methyl

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
		piperidyl sebacate			1,2,2,6,6-pentamethyl-4-piperidyl sebacate
17202	—	SILVER CHLORIDE-COATED TITANIUM DIOXIDE	None	Not retrieved	Silver Chloride; Titanium Dioxide
17408	—	Reaction mass of p-t-butylphenyldiphenyl phosphate and bis(p-t-butylphenyl)phenyl phosphate and triphenyl phosphate	None	Not retrieved	p-t-butylphenyldiphenyl phosphate; bis(p-t-butylphenyl)phenyl phosphate; triphenyl phosphate
17473	—	All salts of Al, NH4, Ba, Ca, Co, Cu, Fe, Li, Mg,	None	Not retrieved	Aluminum octylphosphon

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
		Mn, K, Na, and Zn of n-Octylphosphonic acid			ate; Ammonium octylphosphonate; Barium octylphosphonate; Calcium octylphosphonate; Cobalt octylphosphonate; Copper octylphosphonate; Iron octylphosphonate; Lithium

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
					octylphosphonate; Magnesium octylphosphonate; Manganese octylphosphonate; Potassium octylphosphonate; Sodium octylphosphonate; Zinc octylphosphonate
17601	—	L[TPA+EG]$_2$; [TPA+Et]	None	Not retrieved	Tetraphenylarsonium; Ethylene

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
					glycol; Ethanol
17892	—	All salts of Al, NH4, Ba, Ca, Co, Cu, Fe, Li, Mg, Mn, K, Na, and Zn of Glutaric acid	None	Not retrieved	Aluminum glutarate; Ammonium glutarate; Barium glutarate Calcium glutarate; Cobalt glutarate; Copper glutarate; Iron glutarate; Lithium glutarate; Magnesium glutarate; Manganese

Num.	CAS	Name	Retrieval Source	Retrieved Structure	LLM Prediction
					glutarate; Potassium glutarate; Sodium glutarate; Zinc glutarate
17980	—	trimethylolpropane, diester with 2-ethylhexanoic acid and monoester with benzoic acid	None	Not retrieved	Trimethylolpropane; 2-ethylhexanoic acid; benzoic acid
18130	—	End-group tributyl citrate–rosin ester plasticizer	None	Not retrieved	Tributyl Citrate; Rosin Ester

To validate the LLM-based extraction results, each entry was assigned to one of five categories reflecting the relationship between retrieved structures and predicted components:

1) Correct Prediction: The LLM correctly identified the relevant components described in the name. This category contains a total of 2 entries, including the

following IDs: 16377, 16865.

2) Correct Side-Chain Structure but Incorrect Overall Prediction: The LLM captured core structural fragments (e.g., phosphoric-acid units, glycidyl groups), but the predicted list did not match the actual mixture composition. This category contains a total of 4 entries, including the following IDs: 75, 11236, 4053, 10384.
3) Simplified Structure or Omitted Components: The LLM output chemically reasonable esters or acids (e.g., phthalates, fatty-acid esters), but missed one or more documented component. This category contains a total of 3 entries, including the following IDs: 9485, 11552, 2940.
4) Prediction as Reaction Components: The LLM predicted plausible precursors or reactants rather than the final mixture described in the name. This category contains a total of 6 entries, including the following IDs: 303, 10434, 11237, 17196, 17601, 17980.
5) Undeterminable or Not Found: No authoritative structural information could be retrieved from databases due to lack of CAS numbers or insufficient naming detail. This category contains a total of 25 entries, including the following IDs: 40, 197, 1112, 5892, 8186, 8187, 9894, 11211, 11579, 11632, 12681, 12682, 1498, 16542, 16792, 16843, 17013, 17047, 17202, 17473, 17609, 17892, 18130, 17408, and 17176.

3. Details of ML models for predicting chemical toxicity

3.1 Details of training data for ML

The details of the training data used for the ML models targeting the seven toxicity endpoints are provided below.

Table S7. Class distributions for 7 toxicity indicators.

	C	M	R	CMR	STOT_RE	AqTox	RespSens
0 (non-toxic)	4742	4953	4613	4053	4156	3753	5247
1 (toxic）	835	624	964	1524	1421	1824	330

3.2 Details of ML with multiple molecular representations

1) ECFP: ECFP is a circular fingerprint that encodes local structural features of a molecule based on atomic neighborhood information. A commonly used parameter setting includes a radius of 3 and a bit vector length of 2048 bits.
2) RDKit: RDKFP is a path-based fingerprint that encodes all atom paths within a molecule (typically paths containing 1 to 7 atoms), capturing topological features of the molecular structure. Its default bit vector length is also 2048 bits.
3) MACCS: MACCS is a predefined substructure-based fingerprint consisting of 166 bits, where each bit corresponds to the presence or absence of a specific chemical substructure.
4) ECFP+ RDKit+ MACCS+PCA: The three molecular fingerprints, when concatenated, form a 4,262-dimensional vector. Subsequently, PCA was applied to the concatenated feature matrix, and a cumulative explained variance plot was generated to determine the number of principal components required to retain 85%,

95%, and nearly 100% of the total variance. Ultimately, 2,048 principal components were selected, and the resulting reduced feature matrix was used for training the subsequent MLP model.

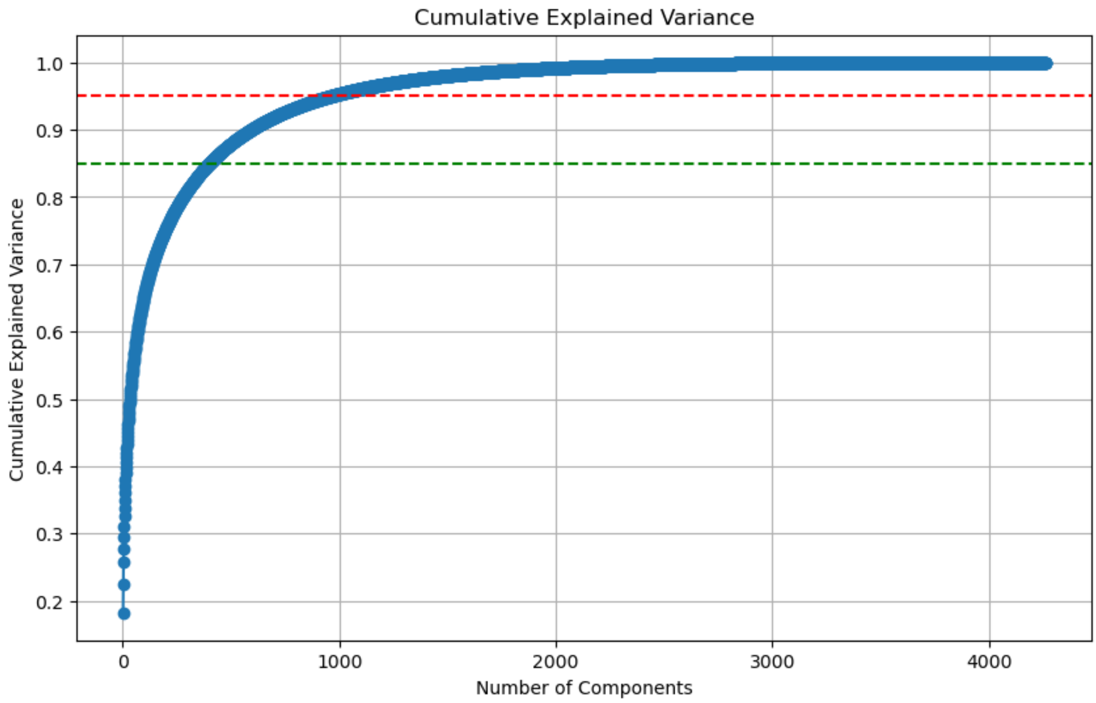

Figure S4. Cumulative explained variance curve of concatenated molecular fingerprints after PCA.

5) GROVER-base: molecular fingerprints generated using the base version of the GROVER model.

6) GROVER-large: molecular fingerprints generated using the large version of the GROVER model.

7) MolCLR-finetune: fine-tuning performed directly on the MolCLR model.

8) MolCLR-feature: hidden-layer embeddings extracted from the MolCLR model and used as molecular fingerprints.

3.3 Results of multi-task learning

ECFP + MACCS fingerprints were used as molecular representations. Single-task: a separate model was trained for each toxicity endpoint. Multi-task: a single model was used to simultaneously predict all endpoints. Selected indicators: to leverage potential inter-task correlations, a subset of related endpoints (C, M, R) was selected for multi-task prediction.

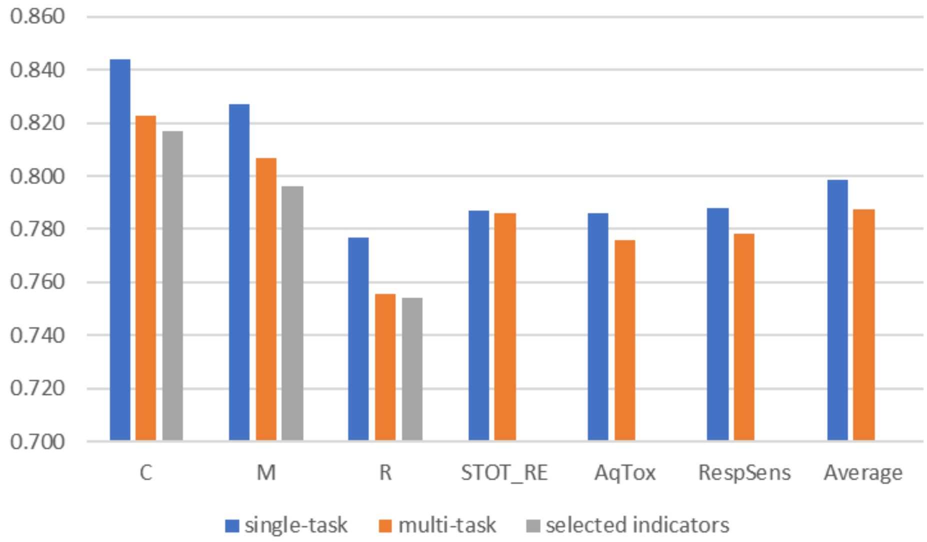

Figure S5. Comparison of multi-task learning and single-task learning performance.

3.4 Results of random hyperparameter optimization

ECFP/MACCS/RDKFP: whether the corresponding molecular fingerprint is used; ECFP_n: length of the generated ECFP fingerprint; ECFP_r: ECFP radius; betas1/betas2: β_1 and β_2 parameters of the Adam optimizer; dp: dropout rate; fps: choice of molecular representation (fingerprints or features generated by GROVER-base or GROVER-large); gamma: γ parameter of the learning rate scheduler; h1/h2: dimensions of the two hidden layers; lr: learning rate; momentum: momentum parameter for the SGD optimizer; op: selected optimizer; schedule: whether to apply a learning rate schedule; step_size: step size parameter of the learning rate scheduler.

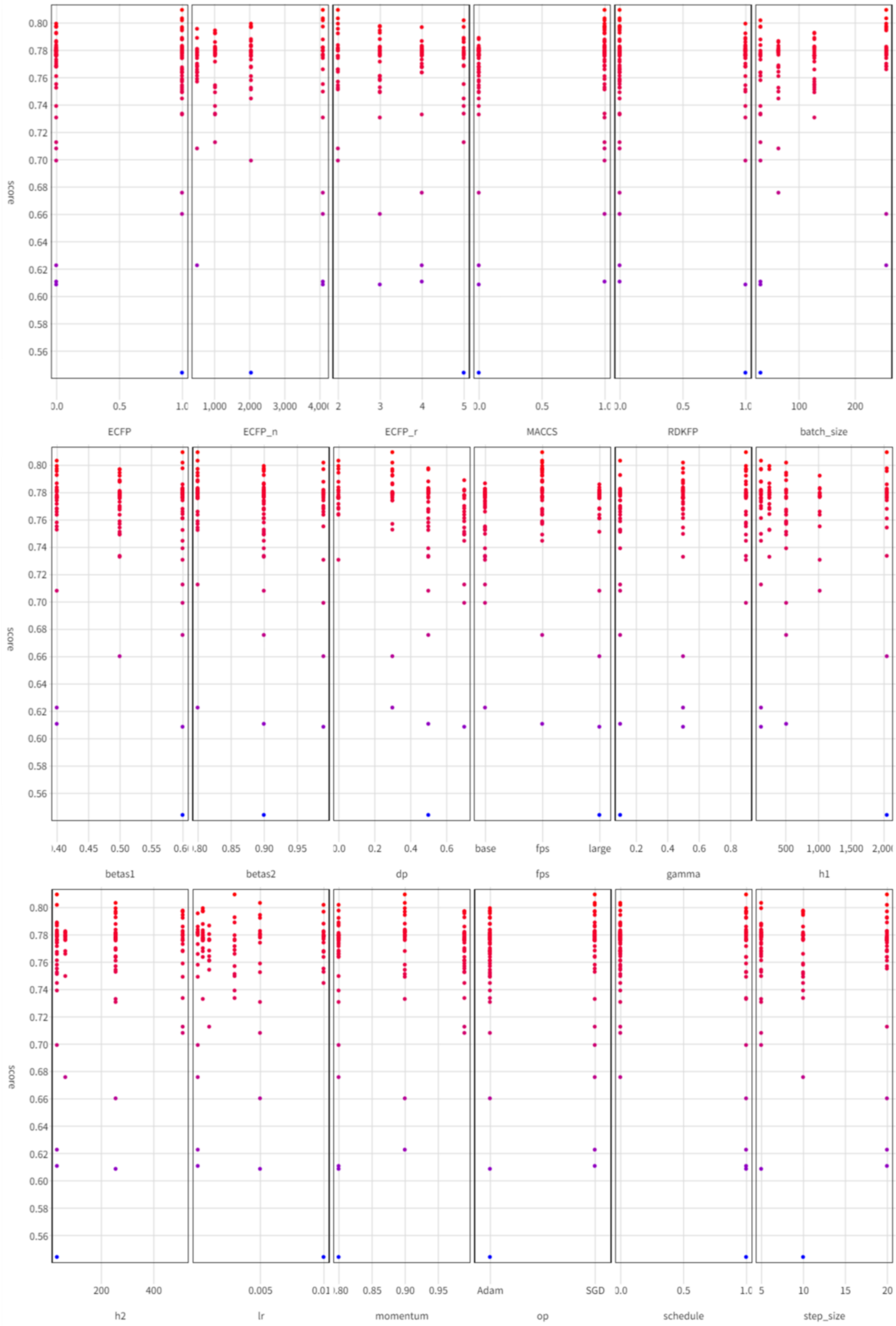

Figure S6. Results of random hyperparameter optimization.

3.5 Results of predictive performance of ML models

The classification accuracies of ML models under different training strategies are summarized below.

Table S8. Toxicity prediction results using different molecular representations.

	CMR	C	M	R	STOT_RE	AqTox	RespSens	Average
ECFP+RDKFP+MACCS+PCA	0.792	0.852	0.825	0.784	0.792	0.787	0.789	0.803
ECFP+RDKFP+MACCS	0.788	0.849	0.822	0.775	0.796	0.790	0.786	0.801
ECFP+MACCS	0.781	0.844	0.827	0.777	0.787	0.786	0.788	0.799
RDKFP+MACCS	0.784	0.844	0.819	0.783	0.790	0.781	0.779	0.797
MACCS	0.772	0.831	0.827	0.768	0.792	0.791	0.777	0.794
ECFP	0.751	0.815	0.800	0.752	0.767	0.761	0.764	0.773
RDKFP	0.773	0.823	0.799	0.769	0.779	0.757	0.766	0.781
GROVER-base	0.770	0.819	0.798	0.756	0.778	0.783	0.770	0.782
GROVER-large	0.771	0.826	0.797	0.763	0.787	0.780	0.781	0.786
MolCLR-finetune	0.764	0.825	0.807	0.760	0.772	0.778	0.787	0.785
MolCLR-feature	0.597	0.664	0.615	0.596	0.649	0.625	0.667	0.630

The following summarizes the validation results of the ML model.

Table S9. Model performance on the final train/validation/test split.

	CMR	C	M	R	STOT_RE	AqTox	RespSens	Average
Valid	0.790	0.852	0.821	0.796	0.795	0.780	0.783	0.802

AUC								
Test AUC	0.762	0.842	0.784	0.729	0.783	0.775	0.840	0.788

3.6 Substructure-level enrichment analysis

In each enrichment plot, the x-axis represents log10(OR), which reflects the strength and direction of the association between a substructure and toxicity (values greater than zero indicate positive enrichment, whereas values below zero indicate negative enrichment). The y-axis denotes the frequency of the substructure in the dataset. For clarity, only substructures exhibiting high-confidence associations are displayed, defined as those with q-values (false discovery rate–adjusted p-value, using the Benjamini–Hochberg correction) below 0.05 and odds ratios greater than 2. To highlight the most informative patterns, four representative substructures are emphasized in each panel by visually marking their occurrences on example molecules. These include the two substructures with the highest odds ratios, which likely represent the strongest toxicity-associated features, as well as the two most frequently occurring substructures among those with odds ratios above 2, which signify commonly appearing features that may contribute to potential toxicity risks. Together, these highlighted substructures provide meaningful mechanistic clues for future toxicological investigations.

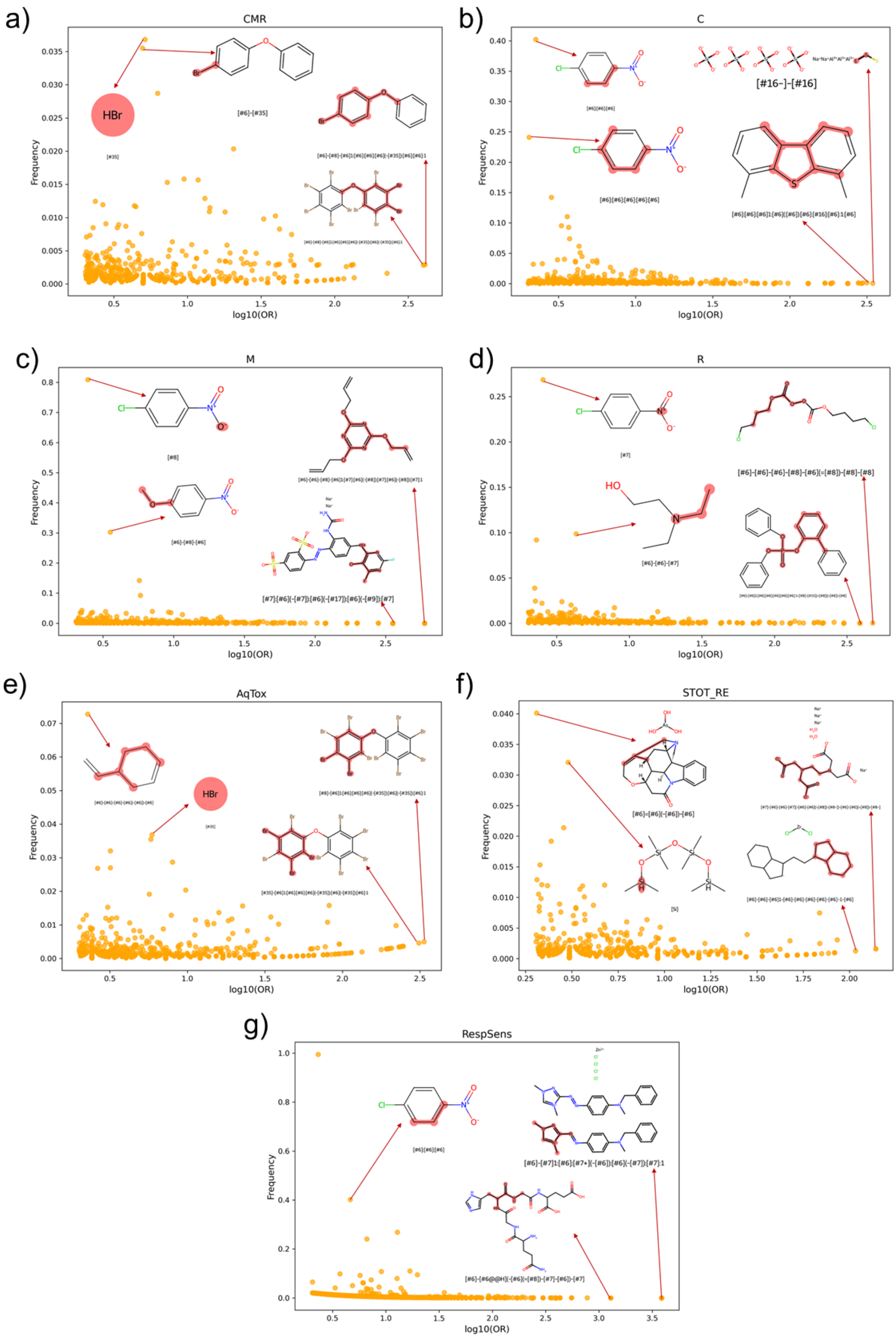

Figure S7. Results of substructure-level enrichment analysis.

3.7 Linking molecular structure-based results to the AOP framework and OECD test guidelines

According to the AOP framework, toxic action can be abstracted as a sequence of molecular structure – molecular initiating event (MIE) – key events (KEs) – adverse outcome (AO) [9,10, 30]. We summarize below the AOP knowledge relevant to the seven toxicity endpoints used in this study, based on the information curated in the AOP-Wiki [11-19].

Table S10. MIE, KE, and AO information of toxicity endpoint (C).

Type	
MIE	Increase, Oxidative DNA damage
KE	Inadequate DNA repair
KE	Increase, DNA strand breaks
AO	Increase, Mutations
AO	Increase, Chromosomal aberrations

Table S11. MIE, KE, and AO information of toxicity endpoint (M).

Type	
MIE	Alkylation, DNA
KE	Inadequate DNA repair
KE	Increase, Mutations
AO	Increase, Heritable mutations in offspring

Table S12. MIE, KE, and AO information of toxicity endpoint (R).

Type	

MIE	Estrogen receptor activation in mammary gland stromal/epithelial cells
KE	Altered transcription in mammary cells
KE	Epigenetic alterations in mammary tissue
KE	Altered cellular differentiation of mammary epithelial cells
KE	Increased collagen deposition in mammary stroma
KE	Increased proliferation of mammary epithelial cells
KE	Altered apoptosis of mammary cells
KE	Altered progesterone receptor signaling
KE	Dedifferentiation of mammary epithelial cells
KE	Disrupted tensional homeostasis in mammary tissue
KE	Desmoplasia in mammary gland stroma
KE	Altered fat pad maturation in the mammary gland
KE	Chronic inflammation in mammary tissue
KE	Increased migration of mammary epithelial/stromal cells
KE	Increased invasion of mammary epithelial cells
KE	Increased mammary gland/breast density
KE	Altered morphogenesis of the mammary gland
KE	Altered hormone sensitivity of the mammary gland
KE	Hyperplasia of mammary epithelium
AO	Enhanced risk for cancer in mammary gland (breast cancer)

Table S13. MIE, KE, and AO information of toxicity endpoint (STOT_RE, OECD Test No. 407).

Type	
MIE	Alkylation, Protein
KE	Increase, Cell injury/death
KE	Tissue resident cell activation

KE	Increased Pro-inflammatory mediators
KE	Activation, Stellate cells
KE	Accumulation, Collagen
AO	N/A, Liver fibrosis

Table S14. MIE, KE, and AO information of toxicity endpoint (STOT_RE, OECD Test No. 408).

Type	
MIE	Inhibition, Bile Salt Export Pump (ABCB11)
KE	Activation of specific nuclear receptors, Transcriptional change
KE	Bile accumulation, Pathological condition
KE	Release, Cytokine
KE	Increase, Inflammation
KE	Increase, Reactive oxygen species
KE	Peptide Oxidation
AO	Cholestasis, Pathology

Table S15. MIE, KE, and AO information of toxicity endpoint (AqTox, AOP312).

Type	
MIE	Acetylcholinesterase (AchE) Inhibition
KE	Acetylcholine accumulation in synapses
KE	Increased Cholinergic Signaling
KE	Impaired coordination and movement
AO	Increased Mortality
AO	Decrease, Population growth rate

Table S16. MIE, KE, and AO information of toxicity endpoint (AqTox, AOP312).

Type	

MIE	Binding of plastoquinone B (QB), PSII antagonism
KE	Decrease, Photosynthesis
KE	Decrease, Coupling of oxidative phosphorylation
KE	Decrease, Adenosine triphosphate pool
KE	Decrease, Cell proliferation
AO	Decrease, Growth

Table S17. MIE, KE, and AO information of toxicity endpoint (RespSens).

Type	
MIE	Covalent Binding, Protein
KE	Increased, secretion of proinflammatory mediators
KE	Activation, Dendritic Cells
KE	Activation/Proliferation, T-cells
AO	Increase, Allergic Respiratory Hypersensitivity Response

Each model endpoint can be mapped to specific traditional toxicological tests. Carcinogenicity corresponds to 18–24 month chronic feeding carcinogenicity studies in rats or mice, in which tumour formation is monitored [20].Mutagenicity corresponds to in vitro mammalian cell micronucleus tests that detect chromosomal aberrations [21]. Reproductive toxicity corresponds to short-term reproductive/developmental toxicity screening studies in rats, where animals are dosed from pre-mating through mating, gestation and early lactation, and parental reproductive performance and early offspring development are observed [22]. STOT_RE_Toxicity_Prediction corresponds to 28-day and 90-day repeated-dose oral toxicity studies in rodents, which jointly evaluate multiple organ systems including liver, kidney, haematology, nervous and immune systems, body weight and pathology [23,24]. AqTox_Toxicity_Prediction corresponds to classical aquatic toxicity tests: 96-h acute fish toxicity tests (e.g. with zebrafish or rainbow trout, recording mortality at 24/48/72/96 h and determining the LC_{50}), 24–48 h Daphnia acute immobilisation tests to determine the EC_{50} for impaired mobility, and 72-h freshwater algae and cyanobacteria growth inhibition tests to determine EC_{50} values [25-27].RespSens_Toxicity_Prediction is informed by air–liquid interface (ALI) models of human bronchial/airway epithelial cells: cells are grown to confluence on porous membranes, the apical medium is removed so that the apical surface is exposed to air while the basolateral surface remains in contact with culture medium, and test substances are applied as aerosols or gases directly onto the cells, followed by determination of IC_{50} values or related cytotoxicity/inflammation endpoints [28,29].

4. Details of fuzzy search method for predicting functional labels

4.1 Results of “conserved sequences”

Among the manually labeled data, only one entry could not be assigned a third-level functional property label suitable for conserved sequence calculation. In addition, major categories without defined subcategories typically exhibit mixed structural features and were therefore excluded from conserved sequence analysis. The conserved sequence results for the remaining third-level functional labels are summarized below.

Table S18. Common sequence of third-level functional labels.

	MACCS common sequence
0	0000000100000000000000000100000000000100011000100011011001101111111101001101010001110100011100101111000011001100101100010111101101001001111100101111101110101011111111111
1	Non
2	Non
3	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001010000100
4	00000000000000000000000000000000000000000000000000000010000000001000000001000000000000000101000001100000110011011001001011010011110110001110111101110101111011111101110
5	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000011100000000000010001011000111100000000110111001000011001110111001110111101110

	MACCS common sequence
6	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100010000000100000000010101000100000001000101100000101001010011110
7	00000000000000000000000000000000100000000000000000100000000000000000000000000010000001100000000100001000010000000010000100000000110000011100000110000011101010000101
8	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000010101000000010010000010000101001110 10
9	000000000000000000000000000000000000000000000000000000000000000010000000000000000000000000100000000000000001000100000000000000000100000000010000000001100001000000000000011010
10	0000000000000000100000100000000000000000000000000000000000010010000000000000000010000100101010000000100000001001100001010000001000000000101101001100010000101011101011011001111
11	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100000000000000000000000000000100000001000000000000
12	000000000000100000000000000000000000000011100010001101000100111100110100000100010011010000100000001000011001100000100000000110011000010000011000010111010000000010010111111
13	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000000001000000000100000000000010000100
14	00000000000000000000000000000000000000000000000000000000000000000000000010000000000000000100000000000000000011000011101000010010110010000010100000110000111101011000100
15	000100
16	00010000100
17	000100

	MACCS common sequence
18	Non
19	00000000000000000000000000000000010000000000000000000000000000001000000010000000000000100000000010000000011001001000000010010100000001010110000100100011101011000100
20	00000000000000000000000000000000000000000000000011001000000000000001000000000010000011100001001000010001100001110100001001111011000011110010011001011110101100011 0
21	000000000000000000000000000010000000000000000001000000001000000001000000010000000000000100000000100010011011001001101010001011111010000110110101111100010101011001110
22	0000000000000000000000000000000000000000000000000000000000000000010000000000000100000000000000000000110000111010000100101100100000100000001101000111010110001 00
23	0000000000000000000000000000000001000110001000110111010011111010010000010000000100100010000111010001100110000011000001000011110011010110011011111011111000111111111 11
24	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001000000100
25	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000110
26	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001001000000100001000100000111010000100101100100000010000001101001011010110001 00
27	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001000000000000000101000000100
28	00000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001010000000000000000101101000100000010011110
29	000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100000000000000000000000000100000000100000000000

	MACCS common sequence
30	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000100
31	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001010000100
32	0000000000000000000000000000000000000000001000000000000000000000000000000000000000000000000000000000010011000010000000000000000000000001000000000000000000000000000000000
33	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100
34	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100000101000100010000000000000000000101000101011000100111100011101010110111110
35	0000000000000000000000000000000000000000000000000010000000000000001000000010100000000000000000000110100010000001000000001000000000000001010001001100010100001010111 1010
36	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001000000000000000000000000100000100000000000000100010100001 00
37	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000001101000000011000000000000000001010000010000 1001010
38	0000000000000000000000000000000000001000000000001000000000000000010000000000010001110000100000001000110001100000001000001100010000000010101000010000000010101101011 11010
39	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100
40	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001010000100
41	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100

	MACCS common sequence
4	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
2	00000000000000000000000000000000000100011000100
4	0000000001000000000000000000000000000000000100010000000001000000000001000001000100000110010100010000111010100001000000101
3	11110010111100011111011001110010111011110100111
4	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
4	00000000000000000000000000000000000000000000100
4	0000000000000000000000000000000000000000000000000000000000000000001000000010000000100000011000000000000000000000100110000100
5	00101111101100010111101111101101111010110111110
4	0000000000000000000000000000000000100011000100011000001001011000001000001100000000000000100000000000000100010100010111010 0
6	00010101110100010001100001111001001010110001 01
4	0000000000000000000000000000000000000000000000000000000100001000000000000000001000000100000000000000000000000000000000000
7	00000000000000000000000000010000000000001011010
4	0000000000000000000000000000000000000011001000010000001001011000001010001000000000100110100100010000101010101101101110100
8	0101010111110001011100000111000010101111100101
4	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
9	00000000000000000000000000000000000000000000100
5	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0	00000000000000000000000000000000000000000000100
5	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
1	00000000000000000000000000000000000000000000100
5	0000000000000000000000000000000000100011000100011111001001111101001010001011011110100001100101111010010001000101000010 11
2	1111101001001011100101101101011101011110111111
5	
3	Non

	MACCS common sequence
54	0000000000000000000000000000000000010001100010001101000100111110100100000101000001001100110010000110011001100010100001000 1101101001001011100101111101011001010111111111
55	0000000001000000000000100000000000010001100011001101100100111100100100000100100111000001110010110101001100110010110001010 0001101001101111000101111101111101011111111111
56	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 0000000000000000000000000000000000000000000100
57	0000000000000000000000000000000000000100000100000000111000000000010000000000010110010000010010111110010001100010010001011 1001101000101011100101111100011001010111111110
58	0000000000000000000000000000000000000000000000000000000000000000000000000010000000000000000000000000000000001100000110100 0010110110010001010100011110110011101011011110
59	Non
60	Non
61	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000010000000000000000000001100001110100 0000010110010000010000000010100010101001000100
62	0000000000000000000000000000000000000000000000000000000000000000000000000010000000000000000000000000000000000000001000000 0000000000000000000000000000100000100001000000
63	0000000000000000000000000000000000000000000000000000000000000000001000000010000000000000000000000000000000000000100000000 0000001000000000000000100000100000001001000110
64	0000000000001000000000000000000000000000000010000100000000000000010000000010000010000110000001000010010101000000000000000 1100100000001110100011001001100000010101111011
65	0000000000000000000000000000000000000011000000010010101001011000001010001000000000100000101000000000001010100000100000100 0001000001110001011000000111000010100010000100

	MACCS common sequence
66	Non
67	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100000000110000000000000010000000000110
68	Non
69	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001000000000000000100000000000001000000110
70	Non
71	Non
72	0000000000000000000000000000000000000000000000000000000001000000000000000011000000010110000010011000100000001111101111101110000011001000011001000001111011011110110 0110
73	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100
74	0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000100
75	Non
76	Non
77	Non

	MACCS common sequence
78	Non
79	0000000001000000000000000000000000000000000010000000000000000000000000100000000000000000000000000000000100000000000000000000010000001000000001000000000000000000001000010
80	Non
81	Non
82	Non

4.2 Similarity scores for third-level functional property labels

Figure S4 displays the structural similarity distribution map for each functional subcategory. Each bubble represents a specific chemical subclass, with the y-axis indicating the consistency in shared structures with existing functional categories (Jaccard similarity). The size of each bubble corresponds to the number of chemicals in that subclass with high structural consistency. For example, classes such as itaconic acid esters, organic secondary stabilizers, and hindered phenolics are concentrated in the similarity range of 0.92–0.98, indicating high similarity with the conserved sequences of known functional labels. In contrast, subclasses with more diverse structures or ambiguous functional boundaries, such as thioether and fatty acid derivatives, show relatively lower similarity and are more widely dispersed.

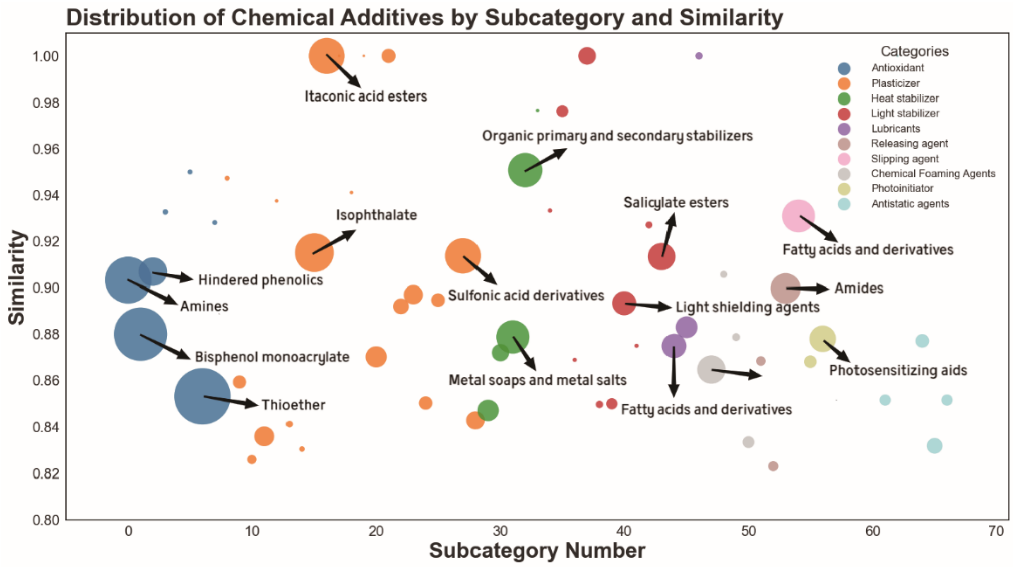

Figure S8. Bubble plot of the distribution of similarity scores among third-level functional labels.

4.3 Manual validation of fuzzy-search functional predictions

In the manual verification results of the fuzzy search, 15 samples were scored 1, 6 samples were scored 0.5, and 19 samples were scored 0.

Table S19. Sampling manual verification results of fuzzy search.

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
0	100-00-5	Benzene, 1-chloro-4-nitro-	Releasing agent	amides	0	Databases such as PubChem/HSDB describe 1-chloro-4-nitrobenzene as an intermediate for dyes, pesticides, rubber chemicals and pharmaceuticals.
1	100-01-6	Benzenamine, 4-nitro-	Releasing agent	amides	0	PubChem/HSDB list 4-nitroaniline as an intermediate for dyes, pharmaceuticals

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
3	100-20-9	1,4-Benzenedicarbonyl dichloride	Antioxidant	Phosphite	0	and pesticides. PubChem and catalogues identify terephthaloyl chloride as a monomer for polyesters and aramids.
7	100-52-7	Benzaldehyde	Plasticizer	terephthalates	0	PubChem and industrial catalogues mainly list benzaldehyde as a fragrance and chemical intermediate, with occasional mention as solvent/plasticizi

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
						ng aid.
9	100-68-5	Benzenamine, 2,4-dichloro-	Antioxidant	Aromatic amines	0	PubChem/Wikipedia: 2,4-dichloroaniline is a dye and agrochemical intermediate.
20	101-77-9	Benzene, 1,1'-methylenebis[4-isocyanato-]	Crosslinking agent	isocyanates	0	Literature and handbooks: 4,4'-methylenedianiline is used to produce MDI and curing agents for polyurethanes.
28	101-84-8	Diphenyl ether	Heat stabilizer	metallic soaps	0	PubChem/HSDB describe diphenyl ether as solvent, heat-transfer fluid and fragrance carrier.

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
36	102-50-1	Benzene, 1-chloro-2,4-dinitro-	Antimicrobial	organic	0.5	PubChem/chemical handbooks: 1-chloro-2,4-dinitrobenzene is mainly an intermediate for dyes and pesticides.
55	104-31-4	Benzaldehyde, 4-methyl-	Antioxidant	Hindered phenolics	0	PubChem/Wikipedia: CAS 104-31-4 corresponds to benzonatate, a cough suppressant.
73	106-47-8	Benzenamine, 4-chloro-	Colorants	azo dyes	0	PubChem/Wikipedia: 4-chloroaniline is used as an intermediate for dyes,

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
						pharmaceuticals and pesticides.
82	108-03-2	Propanamide, N,N-dimethyl-	Anti-sticking agent	amides	0	PubChem lists N,N-dimethylpropanamide as solvent and organic intermediate.
91	108-44-1	Benzenamine, 2,4-dimethyl-	Slipping agent	fatty acids	0	PubChem: 2,4-dimethylaniline is a dye and pesticide intermediate.
103	109-00-2	2-Pyridinecarbonitrile	Polymerization inhibitor	phenols	0	PubChem: 2-cyanopyridine is used to produce nicotinamide and other compounds.
156	115-77-5	Pentaerythritol	Curing agents and	polyols	1	PubChem and coating/resin

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
			curing accelerators			literature describe pentaerythritol as a typical polyol for alkyd and polyurethane resins.
172	119-47-1	Benzene-1,3,5-triol	Antioxidant	polyphenols	1	PubChem/pharmacopoeia: phloroglucinol is a polyhydroxybenzene with antioxidant and pharmacological activity.
203	123-33-1	2,4-Pentanedione	Acid binding agent	chelating	1	PubChem/product sheets: 2,4-pentanedione and analogues are

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
						widely used as metal chelating agents and catalyst ligands.
337	131-11-3	Phthalic acid, diethyl ester	Plasticizer	terephthalates	1	PubChem/handbooks: diethyl (or dimethyl) phthalate is a phthalate plasticizer for cellulose esters etc.
356	133-14-2	2-Naphthalenol, 1-chloro-	Light stabilizer	benzophenones	0	PubChem: 1-chloro-2-naphthol is recorded as an intermediate for dyes and other fine chemicals.
368	133-	9H-	Photoinitia	thioxanthon	1	PubChem/photoi

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
	70-4	Thioxanthen-9-one	tor	es		nitiator literature: thioxanthone is widely used as a UV-curing photoinitiator.
487	141-43-5	Ethanolamine	coupling agent	silanes	0	PubChem/SDS: ethanolamine is used as absorbent, pH adjuster and surfactant raw material.
514	147-14-8	Copper phthalocyanine	Colorants	phthalocyanines	1	PubChem/pigment handbooks: copper phthalocyanine is a common blue pigment.
689	1975-78-	Benzoic acid, 2-	Antioxidant	Aromatic amines	0	Product information: 2-

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
	6	ethoxy-				ethoxybenzoic acid is sold as an organic intermediate.
703	2034-26-6	Silane, methyltrimethoxy-	coupling agent	silanes	1	Silane supplier datasheets: methyltrimethoxysilane is listed as organofunctional silane used as coupling/crosslinking agent.
903	2530-85-0	3-Glycidoxypropyltrimethoxysilane	coupling agent	silanes	1	Technical datasheets: 3-glycidoxypropyltrimethoxysilane (GPTMS) is a typical epoxy-functional silane coupling agent.

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
1102	2991-28-6	1,4-Benzenediol, monoacetate	Heat stabilizer	metallic soaps	0.5	Limited information: 1,4-benzenediol monoacetate is listed as an organic intermediate.
1766	4316-66-3	Propanamide, N-butyl-	Anti-sticking agent	amides	0	Product catalogues list N-butylpropanamide as an organic intermediate.
2205	5528-86-1	Benzamide, N-methyl-	Anti-sticking agent	amides	0	Product catalogues sell N-methylbenzamide as organic intermediate.
3100	6/5/9011	Formaldehyde polymer	Impact Modifiers	methacrylates	0	Information on this “formaldehyde

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
						polymer” entry is scarce and unspecific.
3210	91-68-9	Anthracene, 9-nitro-	Photoinitiator	benzophenones	0.5	Literature and product data: 9-nitroanthracene can be used as photosensitizer or intermediate.
4330	1222-98-6	Bis(hydroxymethyl)propionic acid	Curing agents and curing accelerators	polyols	1	Resin/coating literature: bis(hydroxymethyl)propionic acid is widely used as branching polyol for polyurethanes and polyesters.
6050	68555-77-1	Fatty acids, tall-oil,	Heat stabilizer	metallic soaps	1	Product datasheets: tall-oil fatty acids,

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
		magnesium salts				magnesium salts are considered metallic soaps used for lubrication/stabilization.
6112	68610-51-5	Alcohols, C12-14	Slipping agent	fatty acids	1	Surfactant references: C12–14 alcohols are fatty alcohols used in surfactants and lubricating aids.
7151	125643-61-0	Phosphorous acid, mixed esters	Antioxidant	Phosphite	1	Antioxidant product information: phosphorous acid mixed esters are commonly used as phosphite

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
						secondary antioxidants.
7230	126951-57-7	Triazine derivative	Ultraviolet absorber	triazines	1	UV absorber datasheets: triazine-type UV absorbers are widely used for polymer light stabilization.
11120	25852-47-5	Acrylic acid polymer	Impact Modifiers	methacrylates	1	Polymer literature: poly(acrylic acid) and its esters can act as tackifiers and toughening modifiers.
14020	9003-29-6	Polybutene	Plasticizer	aliphatic esters	0.5	Supplier datasheets: polybutene is often used as

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
						plasticizer, viscosity modifier and tackifier.
15440	9002-88-4	Polyethylene	Nucleator	talc	0	Polymer handbooks: polyethylene is one of the most common commodity plastics.
17533	131545-97-2	Polyamide resin	Releasing agent	amides	0.5	Resin supplier information: polyamide resins are typically used as film-forming resins or binders in coatings and inks.
18952	7732	Water	Chemical	water-based	1	Polyurethane

Num	CAS	Name	Predicted category	Predicted subcategory	Score	Evidence source summary
	-18-5		Foaming Agents			foaming literature: water reacts with isocyanates to generate CO_2 and acts as chemical blowing agent.
19300	9003-55-8	Styrene-butadiene copolymer	Anti-sticking agent	elastomers	0.5	Rubber/plastics handbooks: styrene–butadiene copolymer is an elastomer used in tyres and modified asphalts.

5. Details of exact search method for predicting functional labels

5.1 Details of SMARTS pattern

The SMARTS patterns applied to each plasticizer subclass supported by exact search is provided below.

1) Phthalate:

   First, the SMARTS pattern for phthalates is defined as `c1cc(C(=O)O*)c(C(=O)O*)cc1`, which represents a benzene ring with two ester groups (C(=O)O*) attached to the carbon atoms of the ring, consistent with the characteristic structure of phthalates. Then, the function uses the `Chem.MolFromSmarts()` method to convert the SMARTS pattern into a molecular pattern object and checks whether the input molecule contains the feature structure using `mol.HasSubstructMatch()`. If a match is found, it returns True; otherwise, it returns False. If an error occurs during execution, the function catches the exception and returns False, ensuring the stability of the program.

2) Terephthalate:

   First, the function checks whether the input molecule is a trimellitate ester (by calling the is_Trimellitate_ester function). If it is, the function returns False, excluding this structure. Then, a SMARTS pattern "c1cc(C(=O)O*)ccc1C(=O)O*" is defined to match the terephthalate ester structure (1,4-benzenedicarboxylate diester). This pattern represents a benzene ring where two ester groups (C(=O)O*) are attached at the 1,4 positions, replacing the hydrogen atoms of the benzene ring. The function checks whether the input molecule matches this terephthalate ester structure. If a match is found, it returns True; otherwise, it returns False. If any errors occur during execution, the function returns False.

3) Isophthalate:

First, the function checks whether the input molecule is a trimellitate ester (by calling the is_Trimellitate_ester function). If it is, the function returns False, excluding this structure. Then, a SMARTS pattern "c1ccc(C(=O)O*)cc1C(=O)O*" is defined to match the isophthalate ester structure (1,3-benzenedicarboxylate diester). This pattern represents a benzene ring where two ester groups (C(=O)O*) are attached at the 1,3 positions, replacing the hydrogen atoms of the benzene ring. The function checks whether the input molecule matches this isophthalate ester structure. If a match is found, it returns True; otherwise, it returns False. If any errors occur during execution, the function returns False.

4) Adipic Acid Esters:

First, a SMARTS pattern "O=C(O*)CCCCC(=O)O*" is defined to match the adipate backbone structure. This pattern represents a molecular structure with two ester groups (C(=O)O) attached to a five-carbon chain. The function then checks if the input molecule matches this basic structure. If it does not match, it returns False. Next, the function further excludes molecules containing closed-ring ester structures. A SMARTS pattern "C(=O)O" is defined to match ester groups, and the function checks whether the ester carbon is involved in a ring structure. If the ester carbon is part of a ring, it returns False. The function then continues by matching the ester groups and counting the number of ester groups attached to the ring. If at least two ester groups are attached to the ring, the function returns False. Finally, if all conditions are met, the function returns True, indicating that the molecule is a valid adipic acid ester. If any errors occur during the process, the function returns False.

5) Azelaic Acid Esters:

First, a SMARTS pattern "O=C(O*)CCCCCCCC(=O)O*" is defined to match the structure of azelaic acid esters. This pattern represents a molecular structure with two ester groups (C(=O)O) attached to an eight-carbon chain. The function then checks if the input molecule matches this structure. If it does not match, it returns False. Next, the function excludes molecules containing cyclic esters or alicyclic backbones. It checks for the presence of ring atoms in the molecule, and if a ring structure is detected, it returns False. If the input molecule matches the basic structure of azelaic acid esters and does not contain a ring structure, the function returns True, indicating that the molecule is a valid azelaic acid ester. If any errors occur during the process, the function returns False.

6) Fumaric Acid Esters:

First, the function checks if the input molecule contains more than 8 atoms, as the basic structure of fumaric acid involves 8 atoms. If the number of atoms is less than or equal to 8, the function immediately returns False. Next, a SMARTS pattern "OC(=O)\C=C\C(=O)O" is defined to match the structure of fumaric acid esters. This pattern represents a structure with two ester groups (C(=O)O) and a cis double bond (\C=C) characteristic of fumaric acid. The function uses useChirality=True to enforce consideration of stereochemistry, including cis/trans isomerism of the double bond, and checks whether the input molecule matches the fumaric acid ester structure. If the molecule matches this structure, it returns True; otherwise, it returns False. If any errors occur during execution, the function returns False.

7) Citric Acid Esters:

First, a SMARTS pattern "OC(=O)CC(O)(C(=O)O*)C(C(=O)O*)" is defined to match the structure of citric acid esters or partially esterified structures. This pattern represents a citric acid molecule with three ester groups (C(=O)O) and one hydroxyl group (OH). The function then checks if the input molecule matches this basic structure. If it matches, the function returns True, indicating that the molecule is a citric acid ester. If it does not match, the function returns False. If any errors occur during execution, the function returns False.

8) Trimellitate:

First, a SMARTS pattern "c1c(C(=O)O*)ccc(C(=O)O*)c1C(=O)O*" is defined to match the structure of trimellitate esters. This pattern represents a benzene ring with three ester groups (C(=O)O*) attached at the 1, 2, and 4 positions, replacing the hydrogen atoms of the benzene ring, which is characteristic of trimellitate esters. The function then checks whether the input molecule contains this structure. If a match is found, it returns True, indicating that the molecule is a trimellitate ester. If no match is found, the function returns False. If any errors occur during execution, the function returns False.

9) Itaconic Acid Esters:

First, a SMARTS pattern "O=C(O*)CC(=C)(C(=O)O*)" is defined to match the structure of itaconic acid esters. This pattern represents an itaconic acid ester molecule containing two ester groups (C(=O)O) and a double bond (C=C). The function then checks whether the input molecule matches this structure. If a match is found, it returns True, indicating that the molecule is an itaconic acid ester. If no match is found, the function returns False. If any errors occur during execution, the function returns False.

10) Maleic Acid Esters:

First, the function checks whether the input molecule contains more than 8 atoms. If the number of atoms is less than or equal to 8, it directly returns False. Next, a SMARTS pattern "OC(=O)\\C=C/C(=O)O" is defined to strictly match the structure of cis-maleic acid esters. This pattern represents a maleic acid structure with two ester groups (C(=O)O) and a cis double bond (\C=C). The function uses useChirality=True to enforce consideration of stereochemistry, including cis/trans isomerism of the double bond, and checks whether the input molecule matches this cis-maleic acid ester structure. If the molecule matches the structure, it returns True; otherwise, it returns False. If any errors occur during execution, the function returns False.

11) Oleate:

First, a SMARTS pattern "CCCCCCCC/C=C\CCCCCCCC(=O)O*" is defined to match the esterified form of the oleic acid group. This pattern represents a structure containing 18 carbon atoms, a double bond (C9=C10), and an ester group

(C(=O)O*) that is characteristic of oleic acid. The function then checks whether the input molecule matches this structure. The useChirality=True option ensures that the chirality of the molecule is considered. If the molecule contains this oleate ester structure, it returns True; otherwise, it returns False. If any errors occur during execution, the function returns False.

12) Sebacic Acid Esters,

First, a SMARTS pattern "O=C(O*)CCCCCCCCC(=O)O*" is defined to match the structure of sebacic acid esters. This pattern represents a molecule containing two ester groups (C(=O)O) attached to a nine-carbon chain. The function then checks whether the input molecule matches this basic structure. If it does not match, it returns False. Next, the function excludes molecules containing cyclic esters or alicyclic backbones. It checks for the presence of ring atoms in the molecule, and if a ring structure is detected, it returns False. If the input molecule matches the basic structure of sebacic acid esters and does not contain a ring structure, the function returns True, indicating that the molecule is a valid sebacic acid ester. If any errors occur during execution, the function returns False.

13) Epoxy Derivatives

First, a SMARTS pattern "C1OC1" is defined to match the epoxide ring structure, which represents a three-membered ring containing an oxygen atom. The function checks if the input molecule contains this epoxide group structure. If it does, the function returns True. Next, another SMARTS pattern "CCCC(=O)O" is defined to match long-chain ester groups, representing a structure with an ester group

(C(=O)O) attached to a long carbon chain. The function checks if the input molecule contains this ester group structure. If it does, the function returns True, as the long carbon chain increases the molecular volume, weakens intermolecular forces, reduces material hardness, and improves flexibility. Finally, the function checks if the molecule contains both the epoxide group and the long-chain ester group. If both are present, the function returns True, indicating that the molecule is an epoxy derivative. If either structure is missing, it returns False. If any errors occur during execution, the function returns False.

5.2 Results of the exact search method

The summary of the number of manually labeled data for plasticizers supported by the exact search method and the number of chemicals matched is as follows:

Table S20. Toxicity prediction results using different molecular representations.

Subcategories	Manually labeled data	Matched data
Plasticizer_Phthalate	33	111
Plasticizer_Terephthalate	1	52
Plasticizer_Isophthalate	1	20
Plasticizer_Adipic acid esters	12	53
Plasticizer_Azelaic acid esters	2	8
Plasticizer_Fumaric acid esters	2	17
Plasticizer_Citric acid esters	5	11
Plasticizer_Trimellitate	4	13

Subcategories	Manually labeled data	Matched data
Plasticizer_Itaconic acid esters	2	3
Plasticizer_Maleic acid esters	4	22
Plasticizer_Oleate	5	55
Plasticizer_Sebacic acid esters	7	6
Plasticizer_Epoxy derivatives	5	24

The exact search results for each plasticizer subclass supported by the exact search method, showing the top 50 matched molecules per category, are presented below.

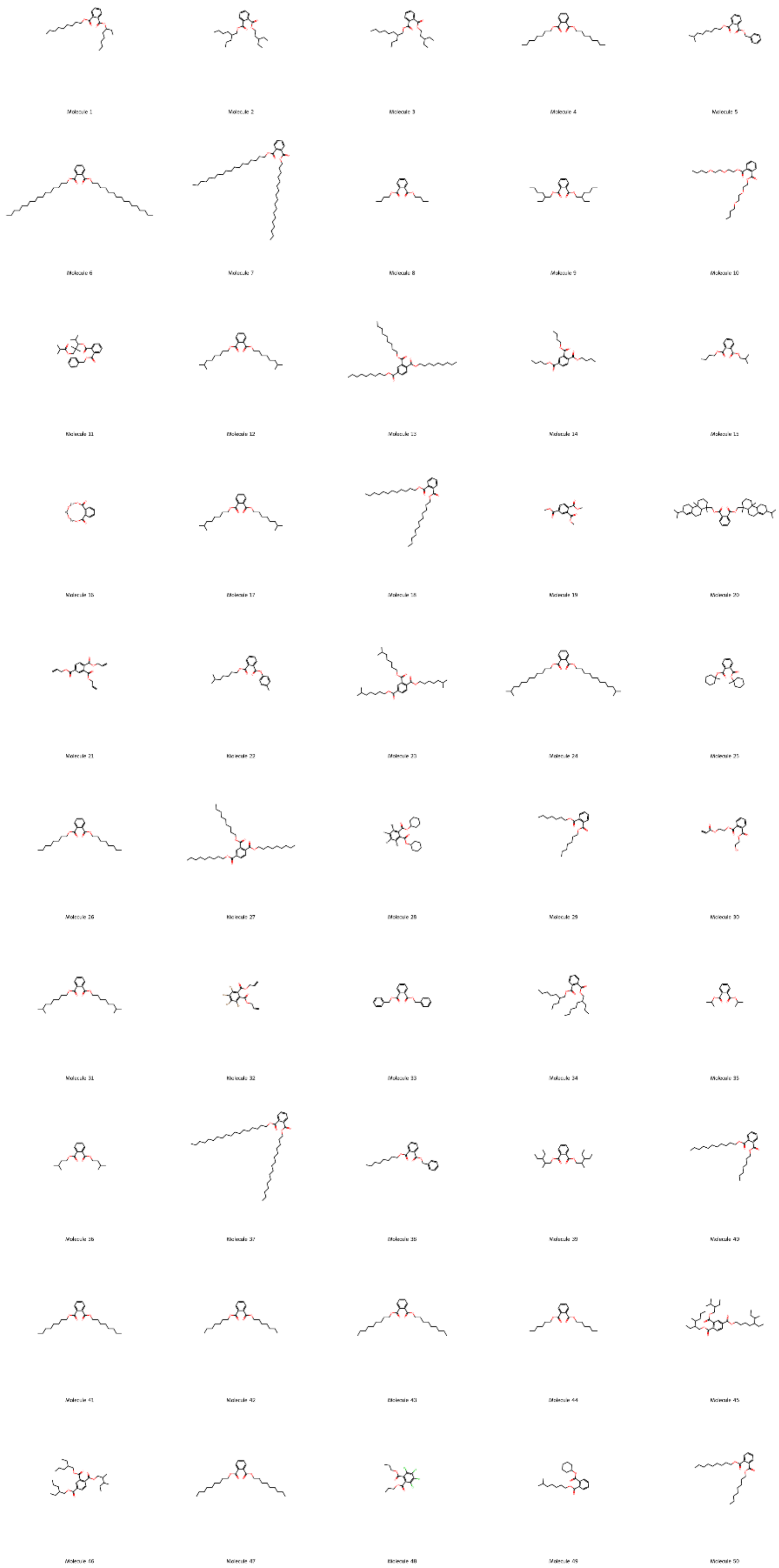

Figure S9. The results of the exact search method for matching phthalate plasticizers are presented below.

Figure S10. The results of the exact search method for matching terephthalate plasticizers are presented below.

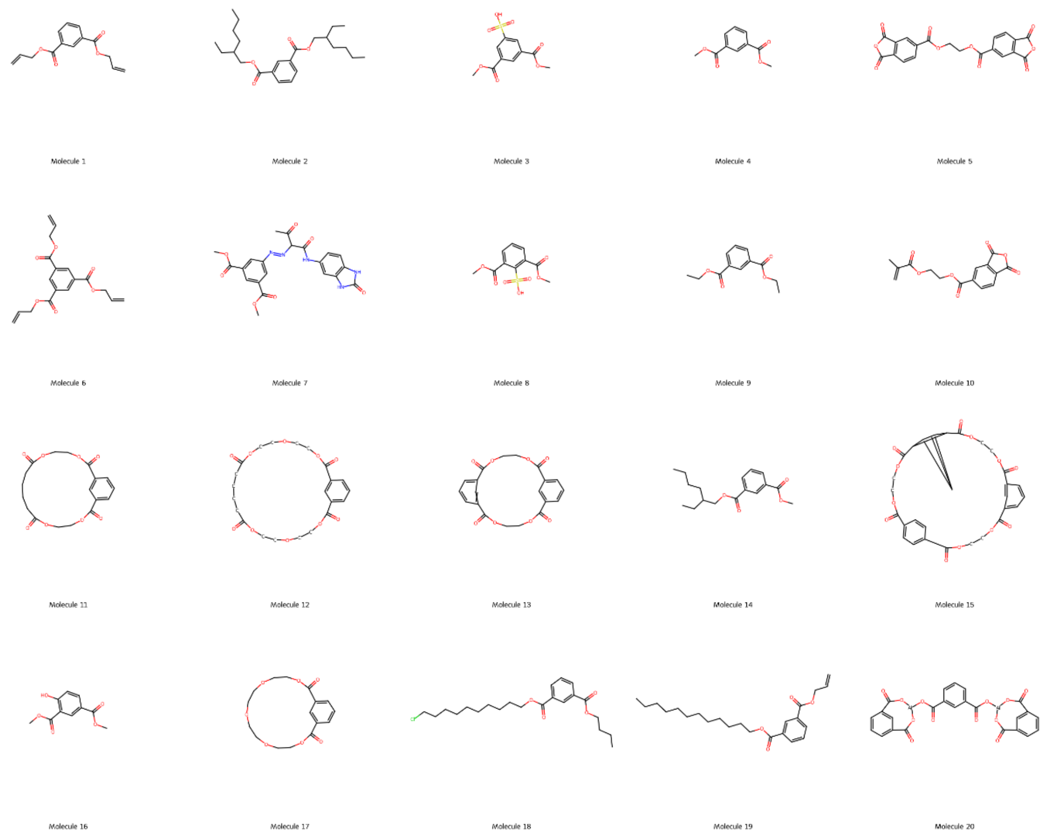

Figure S11. The results of the exact search method for matching isophthalate plasticizers are presented below.

Figure S12. The results of the exact search method for matching adipic acid esters plasticizers are presented below.

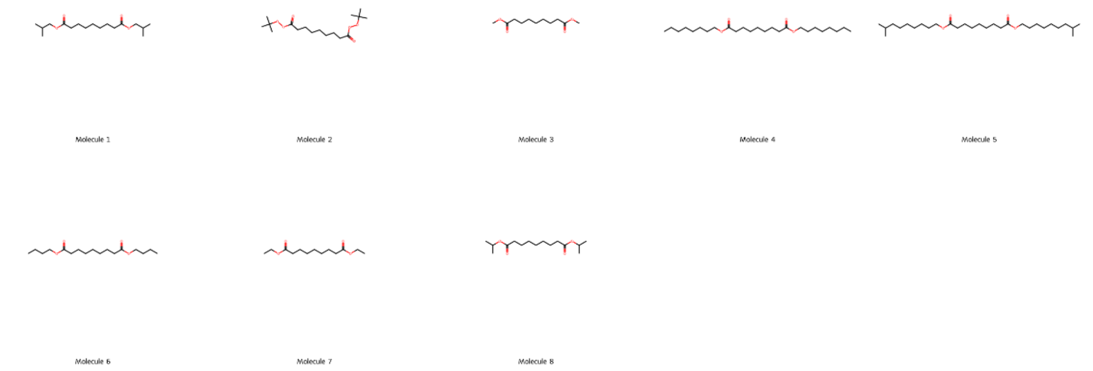

Figure S13. The results of the exact search method for matching azelaic acid esters plasticizers are presented below.

Figure S14. The results of the exact search method for matching fumaric acid esters plasticizers are presented below.

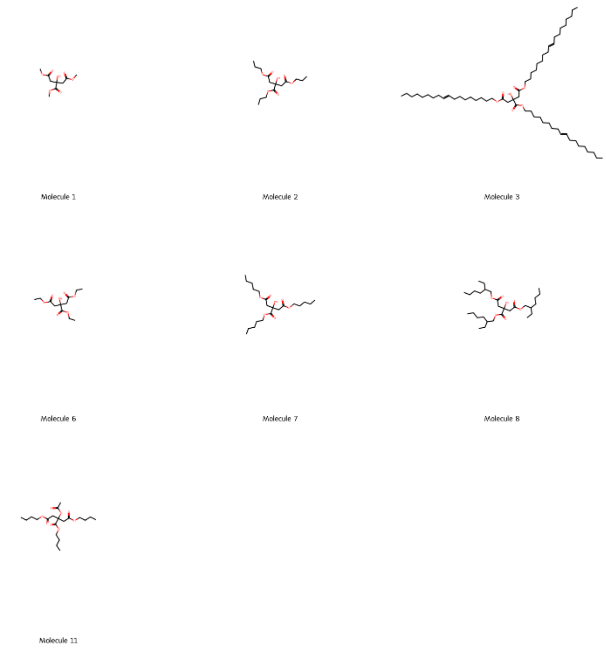

Figure S15. The results of the exact search method for matching citric acid esters plasticizers are presented below.

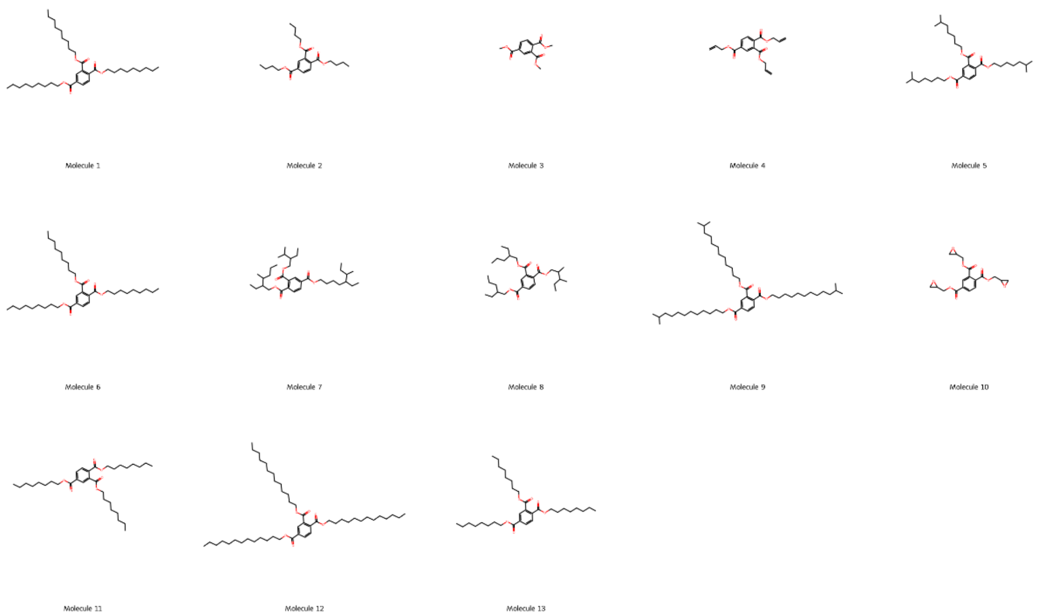

Figure S16. The results of the exact search method for matching trimellitate plasticizers are presented below.

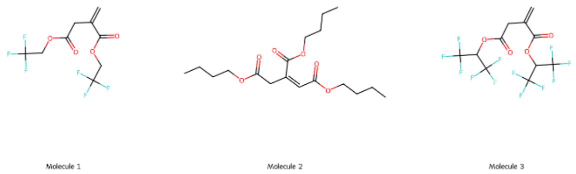

Figure S17. The results of the exact search method for matching itaconic acid esters plasticizers are presented below.

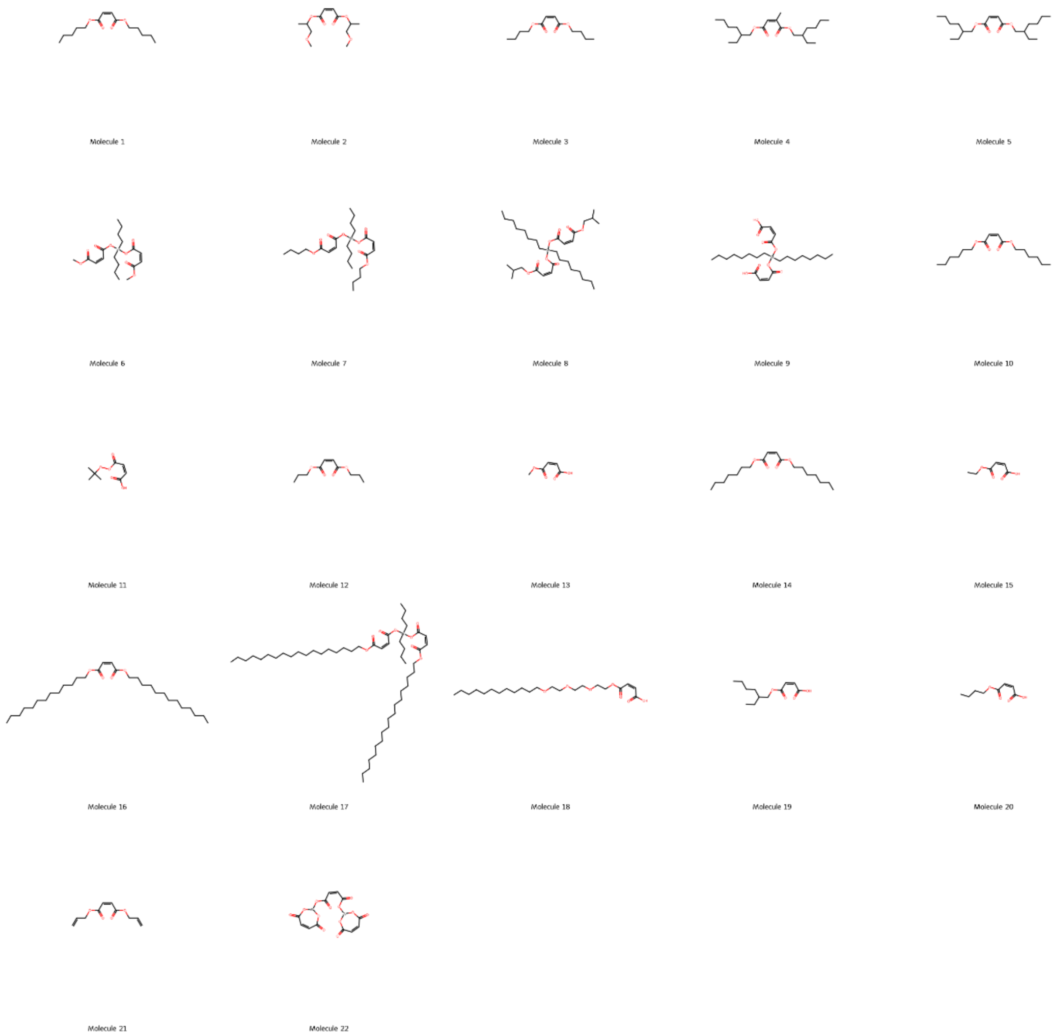

Figure S18. The results of the exact search method for matching maleic acid esters plasticizers are presented below.

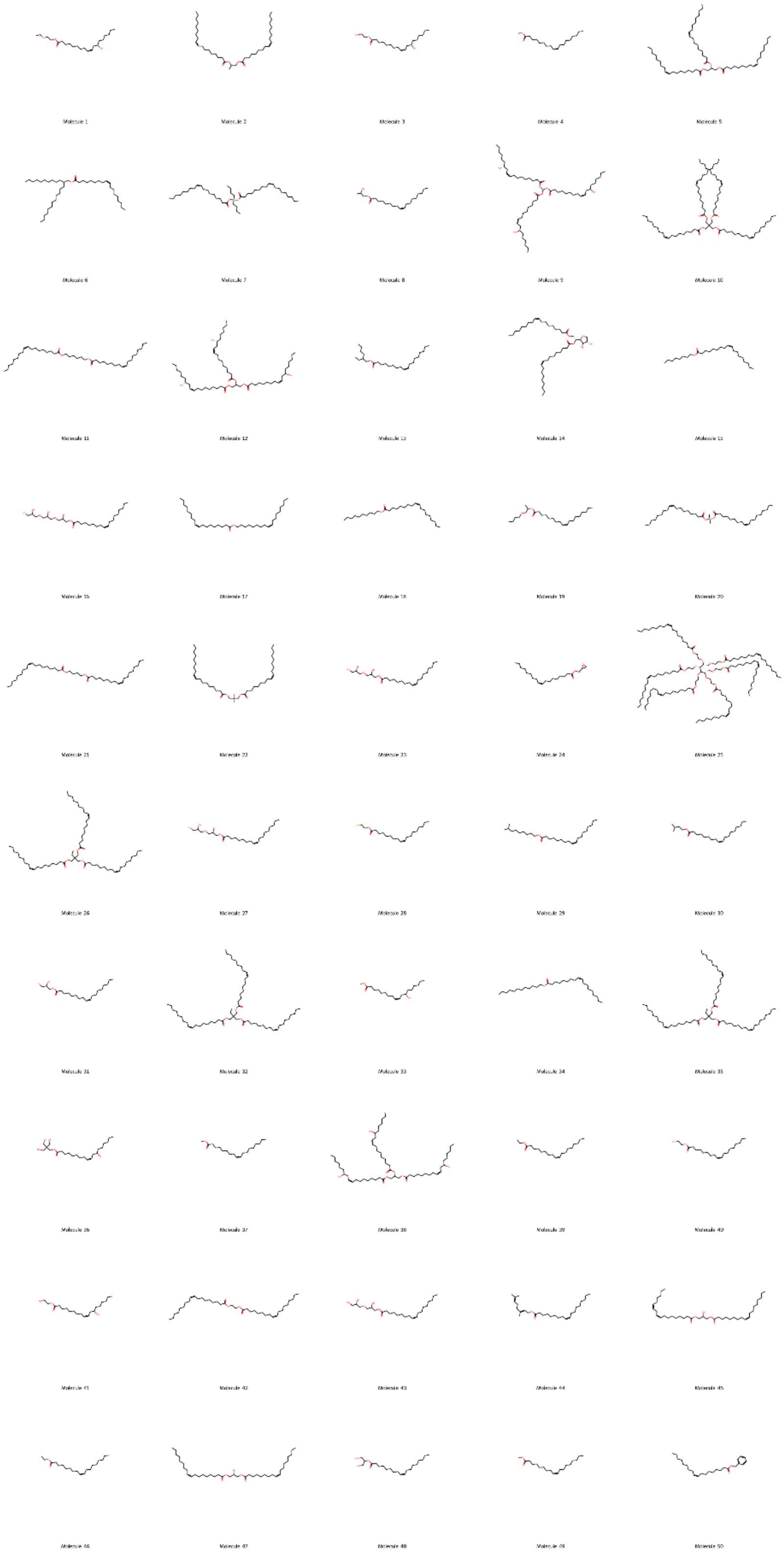

Figure S19. The results of the exact search method for matching oleate plasticizers are presented below.

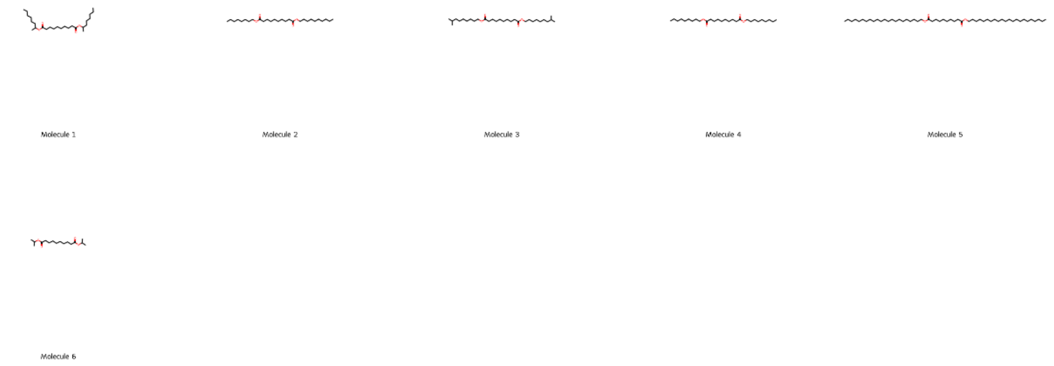

Figure S20. The results of the exact search method for matching sebacic acid esters plasticizers are presented below.

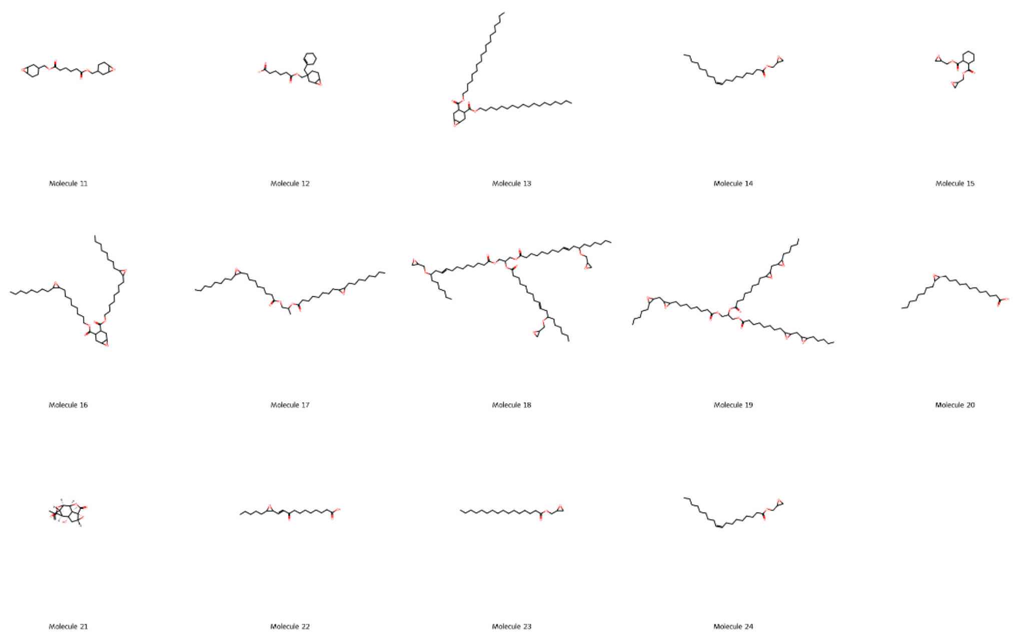

Figure S21. The results of the exact search method for matching epoxy derivatives plasticizers are presented below.

5.3 Results of the exact search method

In the manual validation results of the exact search, the structure scores of 25 samples are 1, the plasticizer classification scores of 13 samples are 1. In the comprehensive scores, 12 samples are scored 1, 13 samples are scored 0.5, and 1 sample is scored 0.

Table S21. Sampling manual verification results of exact search.

Sub-category	CAS	Name	Structural Score	Plasticizer Function Score	Score
Phthalate esters	13988-26-6	Cyclic DEG-PA	1	1	1
Phthalate esters	16883-83-3	1,2-Benzenedicarboxylic acid, 1-[2,2-dimethyl-1-(1-methylethyl)-3-(2-methyl-1-oxopropoxy)propyl] 2-(phenylmethyl) ester	1	1	1

Adipic acid esters	4074-90-2	Hexanedioic acid, 1,6-diethenyl ester	1	0	0.5
Adipic acid esters	—	Adipic acid, di(2-decyl) ester	1	1	1
Sebacic acid esters	10340-41-7	Decanedioic acid, 1,10-bis(1-methylheptyl) ester	1	0	0.5
Sebacic acid esters	124403-19-6	Decanedioic acid, 1-decyl 10-octyl ester	1	0	0.5
Citric acid esters	4552-00-5	Ethyl citrate	1	1	1
Citric acid esters	1587-20-8	2-hydroxy-1,2,3-propanetricarboxylic acid, 1,2,3-trimethyl ester	1	1	1

Epoxy derivatives	26761-45-5	Neodecanoic acid, 2-oxiranylmethyl ester	1	0	0.5
Epoxy derivatives	67860-05-3	2-Oxiraneoctanoic acid, 3-octyl-, 2,2'-(1-methyl-1,2-ethanediyl) ester	1	0	0.5
Oleate	57675-44-2		1	0	0.5
Oleate	126257-84-9	2-decyltetradecyl oleate	1	0	0.5
Trimellitate	94109-09-8	1,2,4-Benzenetricarboxylic acid, 1,2,4-tritridecyl ester	1	1	1
Trimellitate	7237-83-4	1,2,4-Benzenetricarboxylic acid, 1,2,4-tris(2-oxiranylmethyl) ester	1	0	0.5
Maleic acid esters	10099-71-5	2-Butenedioic acid (2Z)-, 1,4-dipentyl ester	1	1	1

Maleic acid esters	31983-42-3	HEPTYL MALEATE	1	1	1
Azelaic acid esters	16580-06-6	Nonanediperoxoic acid, bis(1,1-dimethylethyl) ester	1	0	0.5
Azelaic acid esters	2917-73-9	Nonanedioic acid, 1,9-dibutyl ester	1	1	1
Fumaric acid esters	10341-03-4	2-Butenedioic acid (2E)-, 1,4-ditetradecyl ester	1	1	1
Fumaric acid esters	14595-35-8	PROPYL FUMARATE	1	0	0.5
Itaconic acid esters	104534-96-5	Butanedioic acid, 2-methylene-, 1,4-bis(2,2,2-trifluoroethyl) ester	1	1	0.5

Itaconic acid esters	7568-58-3	1-Propene-1,2,3-tricarboxylic acid, 1,2,3-tributyl ester	1	1	1
Isophthalate	1087-21-4	1,3-Benzenedicarboxylic acid, 1,3-di-2-propen-1-yl ester	1	1	1
Isophthalate	—	Isophthalic acid, butyl 10-chlorodecyl ester	1	0	0.5
Terephthalate	81-30-1	[2]Benzopyrano[6,5,4-def][2]benzopyran-1,3,6,8-tetrone	0	0	0
Terephthalate	34298-51-6	L[TPA+EG]$_4$+EG	1	0	0.5

6. Details of Discussion

6.1 Results of functional–toxicological relationship

1) Direct statistics results

The average toxicity statistics for all chemicals under each third-level functional label (considering only manually labeled chemicals) are presented below.

Table S22. The average toxicity of chemicals corresponding to each third-level functional label

Num.	Sum	CMR	STOT_RE	AqTox	RespSens
C00	0.94	0.00	0.03	0.91	0.00
C01	0.00	0.00	0.00	0.00	0.00
C02	0.00	0.00	0.00	0.00	0.00
C03	0.08	0.00	0.00	0.00	0.00
C04	0.00	0.00	0.00	0.00	0.00
C05	0.00	0.00	0.00	0.00	0.00
C06	0.00	0.00	0.00	0.00	0.00
C07	0.25	0.00	0.00	0.25	0.00
C08	0.50	0.50	0.00	0.00	0.00
C09	0.00	0.00	0.00	0.00	0.00
C10	0.50	0.25	0.00	0.00	0.00
C11	0.00	0.00	0.00	0.00	0.00

Num.	Sum	CMR	STOT_RE	AqTox	RespSens
C12	0.00	0.00	0.00	0.00	0.00
C13	0.00	0.00	0.00	0.00	0.00
C14	0.67	0.00	0.00	0.67	0.00
C15	0.12	0.00	0.00	0.12	0.00
C16	0.60	0.20	0.20	0.20	0.00
C17	0.50	0.42	0.00	0.00	0.00
C18	1.00	0.00	0.00	1.00	0.00
C19	0.50	0.00	0.50	0.00	0.00
C20	0.46	0.13	0.21	0.13	0.00
C21	0.40	0.00	0.13	0.00	0.13
C22	0.00	0.00	0.00	0.00	0.00
C23	0.07	0.00	0.00	0.07	0.00
C24	0.00	0.00	0.00	0.00	0.00
C25	0.36	0.07	0.00	0.21	0.00
C26	0.20	0.00	0.10	0.10	0.00
C27	0.17	0.00	0.00	0.17	0.00
C28	0.00	0.00	0.00	0.00	0.00
C29	0.00	0.00	0.00	0.00	0.00

Num.	Sum	CMR	STOT_RE	AqTox	RespSens
C30	0.15	0.15	0.00	0.00	0.00
C31	0.57	0.00	0.00	0.57	0.00
C32	0.00	0.00	0.00	0.00	0.00
C33	0.10	0.05	0.00	0.05	0.00
C34	1.67	0.33	0.67	0.33	0.00
C35	3.50	1.00	1.00	1.00	0.00
C36	0.50	0.00	0.00	0.50	0.00
C37	0.00	0.00	0.00	0.00	0.00
C38	0.50	0.25	0.00	0.25	0.00
C39	0.65	0.09	0.07	0.40	0.00
C40	2.55	0.73	0.64	0.55	0.00
C41	0.57	0.33	0.05	0.10	0.00
C42	0.25	0.00	0.00	0.00	0.00
C43	0.10	0.00	0.00	0.10	0.00
C44	1.58	0.36	0.24	0.70	0.00
C45	3.22	0.67	0.67	0.89	0.00
C46	1.00	0.00	0.40	0.60	0.00
C47	0.33	0.00	0.00	0.33	0.00

Num.	Sum	CMR	STOT_RE	AqTox	RespSens
C48	0.00	0.00	0.00	0.00	0.00
C49	0.30	0.20	0.00	0.10	0.00
C50	0.25	0.13	0.13	0.00	0.00
C51	1.57	0.21	0.64	0.43	0.00
C52	0.14	0.07	0.04	0.00	0.00
C53	0.00	0.00	0.00	0.00	0.00
C54	0.92	0.03	0.33	0.50	0.00
C55	0.27	0.00	0.18	0.09	0.00
C56	0.25	0.00	0.08	0.00	0.00
C57	0.18	0.14	0.00	0.05	0.00
C58	3.00	0.50	1.00	1.00	0.00
C59	0.71	0.06	0.35	0.24	0.00
C60	0.93	0.27	0.07	0.20	0.07
C61	0.50	0.25	0.25	0.00	0.00
C62	0.50	0.00	0.00	0.50	0.00
C63	0.00	0.00	0.00	0.00	0.00
C64	1.40	0.60	0.40	0.40	0.00
C65	1.67	0.00	0.67	0.33	0.00

Num.	Sum	CMR	STOT_RE	AqTox	RespSens
C66	3.00	1.00	1.00	0.00	0.00
C67	0.00	0.00	0.00	0.00	0.00
C68	0.00	0.00	0.00	0.00	0.00
C69	0.00	0.00	0.00	0.00	0.00
C70	0.00	0.00	0.00	0.00	0.00
C71	0.08	0.00	0.00	0.08	0.00
C72	0.00	0.00	0.00	0.00	0.00
C73	0.44	0.00	0.22	0.22	0.00
C74	0.30	0.00	0.20	0.10	0.00
C75	0.00	0.00	0.00	0.00	0.00
C76	0.80	0.20	0.20	0.40	0.00
C77	2.00	0.33	0.67	0.67	0.00
C78	2.00	0.00	1.00	1.00	0.00
C79	0.00	0.00	0.00	0.00	0.00
C80	1.50	0.50	1.00	0.00	0.00
C81	0.00	0.00	0.00	0.00	0.00
C82	0.00	0.00	0.00	0.00	0.00

To further illustrate these patterns, we calculated, for each functional class and each toxicity endpoint, the proportion of additives with a positive prediction. Since the toxicity labels are binary, this proportion equals the mean of the 0 or 1 values within each functional class. We arranged these proportions into an 83 by 7 matrix and visualized it as a heatmap.

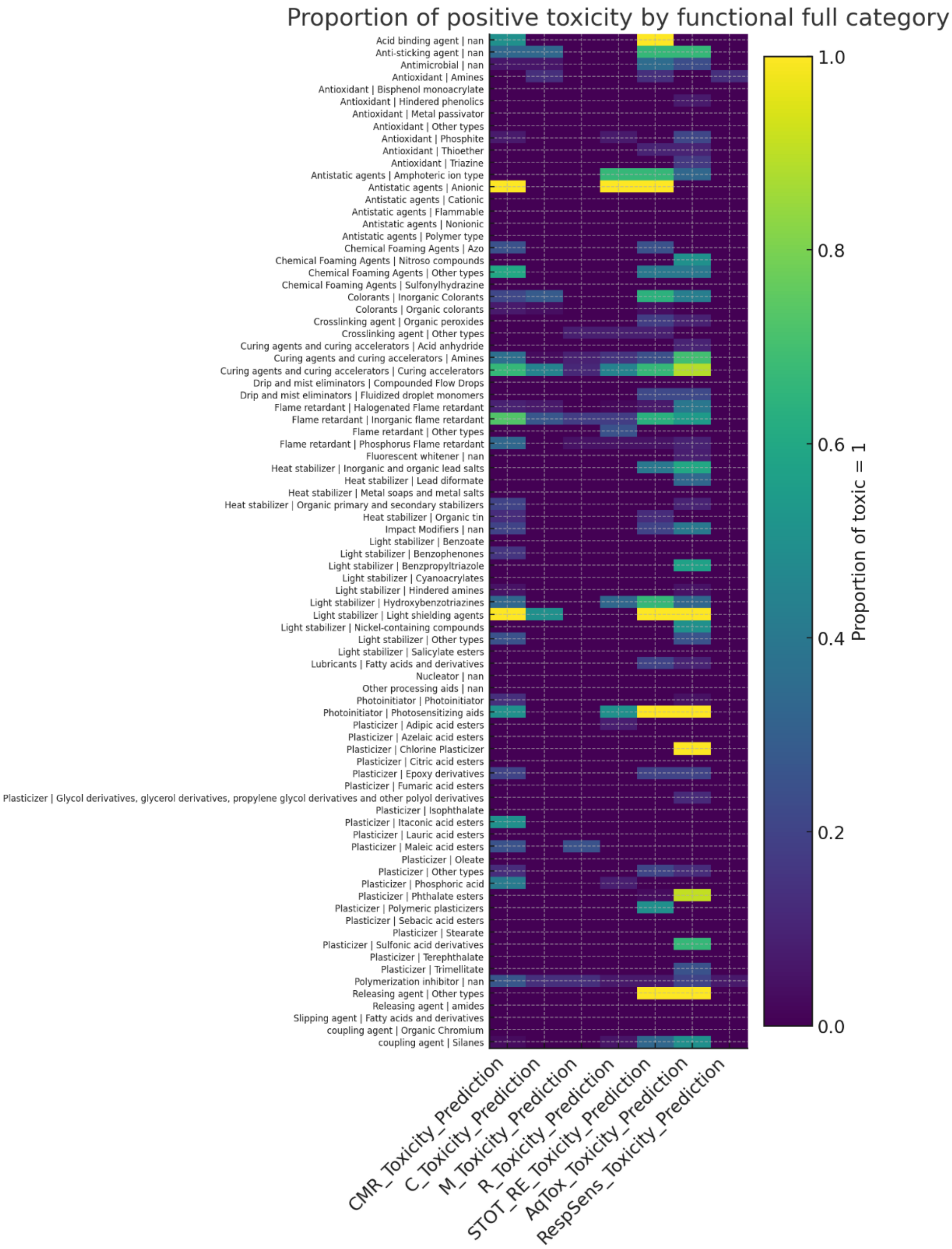

Figure S22. The heatmap of the 83 third-level functional labels and 7 toxicity indicators.

2) Method of Pearson's chi-square test

To test whether the distribution of toxicity labels differs across functional classes, we constructed, for each toxicity endpoint, a contingency table with 83 rows (functional classes) and 2 columns (negative and positive labels). Let O_{ij} denote the observed count in row i and column j of a given contingency table, and let

$$E_{ij} = \frac{(\text{row sum}_i) \times (\text{column sum}_j)}{N}$$

be the expected count under the null hypothesis that functional class and toxicity endpoint are independent, where $N$is the total number of additives. The Pearson chi square statistic is calculated as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

with $r = 83$ and $c = 2$. The degrees of freedom are $\text{df} = (r-1)(c-1) = 82$. For each endpoint, we obtained a p value by comparing the observed $\chi^2$to a chi square distribution with 82 degrees of freedom. A small p value indicates that the observed contingency table is unlikely under the independence assumption and therefore provides evidence of an association between functional class and the toxicity endpoint. In this work we regard p < 0.05 as statistically significant.

Because p values depend on sample size, we additionally quantified the strength of association using Cramer's V, which is a standard effect size for contingency tables. For an $r \times c$table with chi square statistic $\chi^2$and sample size N, Cramer's V is defined as

$$V = \sqrt{\frac{\chi^2}{N \times \min(r-1, c-1)}}.$$

Cramer's V ranges from 0 (no association) to 1 (perfect association). Interpretation thresholds are context dependent, but values around 0.1 are often considered small, around 0.3 moderate and above 0.5 relatively strong.

3) Results of Pearson’s chi-square test

Table S23. Results of Pearson’s chi-square test.

Toxicity endpoint	χ^2	df	p-value	Cramer’s V
CMR_Toxicity_Prediction	223.3297	82	4.999984e-15	0.5442
C_Toxicity_Prediction	160.2123	82	5.445130e-07	0.4610
M_Toxicity_Prediction	85.8802	82	3.630830e-01	0.3375
R_Toxicity_Prediction	179.8895	82	2.715783e-09	0.4884
STOT_RE_Toxicity_Prediction	260.1241	82	2.121680e-20	0.5874
AqTox_Toxicity_Prediction	310.9374	82	2.305561e-28	0.6422
RespSens_Toxicity_Prediction	81.1005	82	5.073262e-01	0.3280

6.2 Illustrative cases of plastic formulation

The formulations of five agricultural plastic films are presented in the following case studies.

Table S24. Different PVC agricultural films use various additives choices in their formulations, calculated using the mass parts counting method [2-8].

Additives	**Formula-tion 1**	**Formula-tion 2**	**Formula-tion 3**	**Formula-tion 4**	**Formula-tion 5**
PVC	100	100	100	100	100
Calcium Carbonate	0	0.5	0	0	0
Carbon Black	0	0	0	0.4	0
Di(2-ethylhexyl) Phthalate	40	25	20	45	40
Di(n-butyl) Phthalate	0	10	10	0	0
Di(2-ethylhexyl) Adipate	7	5	0	0	0
Epoxidized Soybean Oil	3	3	3.5	0	0
Other Plasticizer	0	0	3	5.5	0
Tribasic Lead Sulfate	0	0	2.25	0	0
Glyceryl Monostearate	0	3	0	0	0
Cadmium/Barium Stearate	3	2.5	0	1	1
Dibutyltin Laurate	0	0.5	0	0	0

Ref.

[1] Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. Journal of Medicinal Chemistry 2000, 43 (20), 3714-3717. DOI: 10.1021/jm000942e.

[2] Bing,J.L.;Zhao,J.S.;Bao,Y.Z. Polyvinyl chloride resins and their applications; Chemical Industry Press: China, 2012.

[3] Wang, X.W.; Wang, W.; Liu, Q. Plastic additives and formulation design technologies; Chemical Industry Press: China, 2017.

[4] Lei,C.H.;Xu,R.J.;Dong,Z.H.;Chen,D.H.Plastic materials and additives; China Light Industry Press:China, 2021.

[5] Zuo,J.D.;Luo,C.Y.;Wang,W.G. Plastic Additives and Formulation; Chemical Industry Press: China, 2018.

[6] Li,J.J. Formulation Design of Plastics (3rd Edition); China Light Industry Press:China, 2019.

[7] Gong,L.C.;Zheng,D.;Li,J. Polyvinyl chloride plastic additives and formulation design technology; China Light Industry Press:China, 2006.

[8] Wang,W.G.;Yan,Y.F. Plastic Formulas; Chemical Industry Press: China, 2008.

[9] Cronin, M. T. D.; Richarz, A.-N. Relationship Between Adverse Outcome Pathways and Chemistry-Based In Silico Models to Predict Toxicity. Applied In Vitro Toxicology 2017, 3, 286–297.

[10] Escher, B. I.; Hackermüller, J.; Polte, T.; Scholz, S.; Aigner, A.; Altenburger, R.; et al. From the Exposome to Mechanistic Understanding of Chemical-Induced Adverse Effects. Environment International 2017, 99, 97–106.

[11] AOP-Wiki. Collaborative Adverse Outcome Pathway Wiki; https://aopwiki.org (accessed November 2025).

[12] Cho, E.; Allemang, A.; Audebert, M.; Chauhan, V.; Dertinger, S.; Hendriks, G.; et al. AOP Report: Development of an Adverse Outcome Pathway for Oxidative DNA Damage Leading to Mutations and Chromosomal Aberrations. OECD Series on Adverse Outcome Pathways, No. 29; OECD Publishing: Paris, 2022.

[13] Yauk, C.; Lambert, I.; Marchetti, F.; Douglas, G. Adverse Outcome Pathway on Alkylation of DNA in Male Pre-Meiotic Germ Cells Leading to Heritable Mutations. OECD Series on Adverse Outcome Pathways, No. 3; OECD Publishing: Paris, 2016.

[14] AOP-Wiki. AOP 295: Early-Life Stromal Estrogen Receptor Activation by Endocrine Disrupting Chemicals in the Mammary Gland Leading to Enhanced Cancer Risk; https://aopwiki.org/aops/295 (accessed November 2025).

[15] Landesmann, B. Adverse Outcome Pathway on Protein Alkylation Leading to Liver Fibrosis. OECD Series on Adverse Outcome Pathways, No. 2; OECD Publishing: Paris, 2016.

[16] AOP-Wiki. AOP 27: Bile Salt Export Pump Inhibition Leading to Cholestatic Liver Injury; AOP-Wiki (accessed November 2025).

[17] AOP-Wiki. AOP 312: Acetylcholinesterase Inhibition Leading to Acute Mortality via Impaired Coordination and Movement; AOP-Wiki (accessed November 2025).

[18] AOP-Wiki. AOP 370: Photosystem II Antagonism Leading to Growth Inhibition via Suppression of Photosynthesis; AOP-Wiki (OECD AOP Knowledge Base), https://aopwiki.org/aops/370 (accessed November 2025).

[19] AOP-Wiki. AOP 39: Sensitization of the Respiratory Tract by Low-Molecular-Weight Organic Chemicals; AOP-Wiki (accessed November 2025).

[20] OECD. Guidance Document 116 on the Conduct and Design of Chronic Toxicity and Carcinogenicity Studies, Supporting Test Guidelines 451, 452 and 453, 2nd ed.; OECD Publishing: Paris, 2012.

[21] OECD. Test No. 487: In Vitro Mammalian Cell Micronucleus Test. OECD Guidelines for the Testing of Chemicals, Section 4; OECD Publishing: Paris, 2010.

[22] OECD. Test No. 421: Reproduction/Developmental Toxicity Screening Test. OECD Guidelines for the Testing of Chemicals, Section 4; OECD Publishing: Paris, 2015.

[23] OECD. Test No. 407: Repeated Dose 28-Day Oral Toxicity Study in Rodents. OECD Guidelines for the Testing of Chemicals, Section 4; OECD Publishing: Paris, 2008.

[24] OECD. Test No. 408: Repeated Dose 90-Day Oral Toxicity Study in Rodents. OECD Guidelines for the Testing of Chemicals, Section 4; OECD Publishing: Paris, 2018.

[25] OECD. Test No. 203: Fish, Acute Toxicity Test. OECD Guidelines for the Testing of Chemicals, Section 2; OECD Publishing: Paris, 2019.

[26] OECD. Test No. 202: Daphnia sp. Acute Immobilisation Test. OECD Guidelines

for the Testing of Chemicals, Section 2; OECD Publishing: Paris, 2004.

[27] OECD. Test No. 201: Freshwater Alga and Cyanobacteria, Growth Inhibition Test. OECD Guidelines for the Testing of Chemicals, Section 2; OECD Publishing: Paris, 2011.

[28] Sullivan, K. M.; Enoch, S. J.; Ezendam, J.; Roggen, E. L.; Aleksic, M.; Aspinall, A.; et al. An Adverse Outcome Pathway for Sensitization of the Respiratory Tract by Low-Molecular-Weight Organic Chemicals. Applied In Vitro Toxicology 2017, 3, 213–226.

[29] Upadhyay, S.; Palmberg, L. Air–Liquid Interface: Relevant In Vitro Models for Investigating Air Pollutant-Induced Pulmonary Toxicity. Toxicological Sciences 2018, 164, 21–30.

[30] OECD. OECD QSAR Toolbox; https://qsartoolbox.org (accessed November 2025).