Dataset Description (RAG)

A) Manuscript: 3WR-2025-manuscript.docx

Contains: ToxD4C multi-modal multi-task design (fingerprints + graph + 3D SE(3)-Transformer), Uni-Mol transfer learning on DFT-optimized structures, calibration/uncertainty concepts, SHAP-based interpretation, and docking/validation workflow (ERβ/ERα/AR/PPARγ; cellular assays such as JC-1, CCK-8, reporter pathways).

B) Latest study main paper: 1-s2.0-S0160412025005999-main.pdf

Contains: TBBPA metabolism in HLM/HepG2/MIHA, transformation pathways, and environmental/toxicological implications.

C) Latest study supplementary: 1-s2.0-S0160412025005999-mmc1.docx

Contains: metabolite IDs and structural classes (e.g., sulfation, glucuronidation, methylation, glycosylation, debromination, coupling, substitution), and identification evidence.

D) Toxicity prediction results (31 endpoints): smiles2_toxicity_results.csv

21 molecules × 31 endpoints (26 classification + 5 regression), including Probability/Result/Risk_Level for classification and LD50/LC50/IGC50/BCF-like regressions.

E) Molecular descriptor / DFT-property prediction table: prediction_results_with_smiles.csv

21 molecules × 52 columns, keyed by SMILES, including the following descriptor fields: SMILES, AtomNum, Weight, HOMO, HOMO_number, LUMO, HOMO_LUMO_Gap, ODI_HOMO_1, ODI_HOMO, ODI_LUMO, ODI_LUMO_Add1, ODI_Mean, ODI_Std, Farthest_Distance, Mol_Radius, Mol_Size_Short, Mol_Size_2, Mol_Size_L, Length_Ratio, Len_Div_Diameter, MPP, SDP, Dipole_Moment, Quadrupole_Moment, Octopole_Moment, Volume, Density, ESPmin, ESPmax, Overall_Surface_Area, Pos_Surface_Area, Neg_Surface_Area, Overall_Average, Pos_Average, Neg_Average, Overall_Variance, Nu, Pi, MPI, Nonpolar_Area, Polar_Area, ALIEmin, ALIEmax, ALIE_Ave, ALIE_Var, LEAmin, LEAmax, LEA_Ave, LEA_Var, XLogP, TPSA, Complexity

F) Any provided figures/images (if accessible) that help verify metabolite classes or highlight outliers.

Prompts:

You are a research-grade scientific agent specializing in environmental toxicology, computational chemistry, and explainable ML. Your mission is to explain—mechanistically and causally—WHY the predicted toxicity profiles (31 endpoints) of the newly reported TBBPA transformation/metabolite products look the way they do, by integrating:

(1) the modeling framework + conclusions in my manuscript (ToxD4C + Uni-Mol transfer learning + SHAP + docking/validation logic), and (2) the newest TBBPA metabolism/transformation study (main text + supplementary metabolite IDs / pathways), (3) my toxicity prediction table (31 endpoints), and (4) my molecular descriptor/DFT-property prediction table (quantitative descriptors per SMILES).

You must not treat predictions as ground truth. All statements must be phrased as “model-predicted signals / hypotheses requiring validation”, and you must separate evidence-based claims from speculation.

PRIMARY OUTPUT GOALS

Molecule-level diagnosis: For each molecule, identify the endpoints with HIGH / abnormal risk (especially NR_* nuclear receptor endpoints and SR_* stress-response endpoints),

and explain the likely drivers using a structured causal chain: “Transformation type → structural motif changes → descriptor shifts → SHAP-consistent drivers → plausible receptor/pathway mechanism → testable validation plan”.
Transformation-class level rules: Group the 21 molecules by transformation type (debromination, sulfation, glucuronidation, glycosylation, methylation, coupling/dimerization, substitution, etc.) using the latest study’s metabolite IDs where possible; otherwise infer from substructure rules. For each class, summarize why toxicity signals rise/fall or split across endpoints, and reconcile with the newest study’s pathway logic.
Model credibility & applicability domain: Identify where predictions may be less reliable (e.g., highly ionized conjugates, large polar adducts, high flexibility; potential out-of-domain relative to Tox21 chemical space; single-conformer limitations). Flag likely false positives/negatives and propose counterfactual checks.

MANDATORY WORKFLOW

Step 1 — Extract “interpretation priors” from the manuscript 1.1 Summarize the modeling stack: ToxD4C inputs, dynamic fusion (cross-attention + gating), multi-task endpoints (26 cls + 5 reg), transfer learning on DFT-optimized 3D structures (Uni-Mol). 1.2 Extract the SHAP-driven key drivers and their directional effect on toxicity signals (e.g., XLogP, HOMO–LUMO gap, Quadrupole moment, etc.). Record any threshold-like reference values mentioned in the manuscript (Table/Results), but explicitly state they are dataset/model-specific (not universal toxicology laws). 1.3 Extract the manuscript’s mechanistic validation logic: which endpoints map to which receptors (ERβ/ERα/AR/PPARγ) and how docking/assays were used to support interpretations.
Step 2 — Build a unified analysis table (join toxicity + descriptors by SMILES) 2.1 Load smiles2_toxicity_results.csv. For each molecule, build an “endpoint signature”: - Classification: count HIGH/MEDIUM/LOW; list all HIGH endpoints with probabilities. - Emphasize NR_* and SR_* endpoints (NR-ER/AR/AhR/Aromatase/PPARγ; SR_ARE/HSE/p53/ATAD5/MMP, etc.). 2.2 Load prediction_results_with_smiles.csv and merge by SMILES to attach quantitative descriptors. 2.3 For each endpoint (especially those frequently HIGH), analyze which descriptors co-vary with risk: - Use rank correlations and/or simple monotonic screening (no overfitting; the point is interpretability). - Focus first on descriptors highlighted in the manuscript SHAP narrative (e.g., XLogP, HOMO_LUMO_Gap, Quadrupole_Moment, Polar/Nonpolar area, ALIE/LEA metrics, ESP extremes). 2.4 Produce three Top-5 priority lists: (i) overall hazard (most HIGH / strongest combined signal), (ii) endocrine / nuclear receptor priority (NR_*), (iii) mitochondrial & stress-response priority (SR_MMP, SR_ARE, SR_p53, SR_ATAD5, SR_HSE). For each Top molecule: name the specific endpoints driving its rank and the top 3 descriptor anomalies relative to the cohort.
Step 3 — Metabolite mapping and transformation-class grouping 3.1 From the latest study (main + supplement), map metabolite IDs (e.g., Mxxx) / classes to the SMILES in my tables when possible. 3.2 If IDs are missing in my csv, infer transformation classes using structure rules: - sulfation: sulfate ester / sulfonate-like motifs - glucuronidation: glucuronic acid conjugate motif - glycosylation: sugar ring conjugate - methylation: methoxy substitution / O-methyl ether - debromination: reduced Br count - coupling/dimerization: doubled

aromatic framework / linked rings - substitution/elimination patterns per the newest study 3.3 For each class: compute the class-average “endpoint signature” and class-average descriptor profile, and identify within-class outliers.

Step 4 — Mechanistic “cause chain” explanations (core deliverable) For each priority molecule (and any critical outlier), write a testable, multi-layer explanation that MUST include all four layers:

4.1 Structural transformation layer: Specify how it differs from parent TBBPA or from its nearest class analog: debromination degree, conjugation type (sulfate/glucuronide/sugar), methylation, coupling, etc. Note resulting changes in HBD/HBA, ionization propensity, aromatic exposure, sterics, flexibility.

4.2 Descriptor / physicochemical layer (use the actual descriptor table; do not guess when data exist): Compare its XLogP, TPSA, Weight, Dipole/Quadrupole/Octopole moments, Polar/Nonpolar area, ESPmin/ESPmax, ALIE/LEA, MPI, etc. to cohort medians and to any manuscript SHAP priors/threshold references. Explain how these shifts plausibly affect membrane partitioning, bioavailability, bioaccumulation (BCF), and receptor-binding feasibility.

4.3 Electronic/reactivity layer: Use HOMO, LUMO, HOMO_LUMO_Gap, ESP extremes, ALIE/LEA metrics, and quadrupole/dipole features to argue (qualitatively if needed) about electrophilicity/nucleophilicity, H-bonding/halogen-bond propensity, and noncovalent interaction tendencies that could drive NR or SR endpoints. If a claim is inferential, label it explicitly as a hypothesis.

4.4 Pathway/receptor layer aligned to endpoints: - NR_ER / NR_ER_LBD: relate to phenolic OH patterning, aromatic geometry, halogen effects, and steric fit. Recommend docking to ERβ/ERα if needed, and specify what binding features to check (H-bond anchors, hydrophobic packing, halogen interactions). - NR_PPARγ: relate to hydrophobic bulk + polar headgroup compatibility; discuss how sulfation/glucuronidation might reduce permeability yet still show binding signals (or produce model artifacts) and propose how to differentiate. - SR_MMP: connect to membrane/mitochondrial partitioning (XLogP, nonpolar area), aromatic halogenated scaffolds, and potential uncoupling-like behavior; propose JC-1 validation logic. - SR_ARE/HSE/p53/ATAD5: connect to oxidative stress, proteostasis stress, DNA damage/replication stress; leverage reactivity proxies (Gap, ESP extremes, ALIE/LEA).

You MUST produce an “Explanation Matrix”: One row per priority molecule, columns: [Transformation type] → [Key descriptor shifts] → [Manuscript-SHAP-consistent driver direction] → [High-risk endpoints] → [Most plausible mechanism] → [Validation plan (assay + docking/MD + additional computations)].

Step 5 — Applicability domain + counterf