

Exposure-Route-Driven Four-Stage Computational Toxicology Assessment Paradigm for Cosmetic Ingredient Safety Evaluation

Fangfang Zhou

Guangdong Institute for Drug Control

Dong Guo

g1334524567@163.com

Guangdong Institute for Drug Control

Yanjia Cai

Guangdong Institute for Drug Control

Jingwen Liang

Guangdong Institute for Drug Control

Yangjie Li

Guangdong Institute for Drug Control

Qunyue Wu

Guangdong Institute for Drug Control

Jihui Fang

Guangdong Institute for Drug Control

Research Article

Keywords: Computational toxicology, Exposure route, Next-generation risk assessment, Adverse outcome pathway, Weight of evidence, Cosmetic safety

Posted Date: September 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11055032/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations:

No competing interests reported.

Figures are available in the Supplementary Files section.

Table 1 is available in the Supplementary Files section.

1 **Title page**
2 **Exposure-Route-Driven Four-Stage Computational**
3 **Toxicology Assessment Paradigm for Cosmetic Ingredient**
4 **Safety Evaluation**

5

6 Fangfang Zhou^{1, 3} · Dong Guo^{1, 2, *} · Yanjia Cai^{1,3} · Jingwen Liang¹ · Yangjie Li^{1,}
7 ² · Qunyue Wu^{1, 2} · Jihui Fang^{1, 3*}

8 ¹ NMPA Center for Innovation and Research in Regulatory Science, Guangdong
9 Institute for Drug Control, Guangzhou 510006, China

10 ² Guangdong Provincial Biomedical Science and Technology Collaborative
11 Innovation Center, Guangdong Institute for Drug Control, Guangzhou 510006, China

12 ³ School of Pharmaceutical Sciences, Guangzhou Medical University, Guangzhou
13 511436, China

14

15 Fangfang Zhou and Dong Guo contributed equally to this work.

16

17 *Corresponding author:

18 Dong Guo (g1334524567@163.com),

19 Jihui Fang (jhfayy@163.com)

20 Address:

21 Guangdong Institute for Drug Control, Guangzhou 510006, China

22 Tel./Fax: +86-20-81848153

Abstract

Cosmetic safety assessment has shifted toward next-generation risk assessment (NGRA); although methodological principles for integrating heterogeneous evidence are established, the computational toxicology layer has long been an adjunct to in vitro studies, lacking independent functional boundaries and systematic tool-coordination logic. We reconstruct exposure information from a static parameter into a routing variable and present a “screening–exposure–mechanism–decision” four-stage computational toxicology assessment paradigm, in which molecular-initiating-event-driven, adverse outcome pathway -maturity-moderated three-mode functional mapping defines the correspondence between methods and biological anchors, and weight-of-evidence (WoE) priority and conflict-resolution rules enable transparent integration of heterogeneous evidence. The paradigm’s operability is demonstrated with homosalate as a case study. Exposure routing expanded the assessment object from parent compound to metabolites; loop screening identified a progesterone receptor (PR) interference signal in CYP-mediated oxidative metabolites that was absent in the parent compound but positive in 10/10, demonstrating that parent-based hazard profiles cannot cover metabolite-specific signals and that metabolites must be explicitly included and re-screened. WoE integration resolved the directional conflict for skin sensitization toward a Credible Negative and, on computational evidence, a Clear Positive for the estrogen receptor α pathway, while metabolite PR interference was flagged as Positive with Concern, with in vitro verification recommended under integrated approaches to testing and assessment guidance. The paradigm provides a computation-driven pathway from hazard identification to decision output, positioning the *in silico* layer as a structured contributor within the NGRA evidence chain.

Keywords

Computational toxicology · Exposure route · Next-generation risk assessment · Adverse outcome pathway · Weight of evidence · Cosmetic safety

Abbreviations

47 ADI, Applicability Domain Index; AEP, Aggregate Exposure Pathway; AOP, Adverse Outcome
 48 Pathway; AR, Androgen Receptor; CIR, Cosmetic Ingredient Review; CYP, Cytochrome P450; ECHA,
 49 European Chemicals Agency; EFSA, European Food Safety Authority; ER, Estrogen Receptor;
 50 ER α /ER β , Estrogen Receptor α/β ; GR, Glucocorticoid Receptor; HCM, Hazard Cheminformatics
 51 Module; HMS, Homosalate; HQ, Hazard Quotient; IATA, Integrated Approaches to Testing and
 52 Assessment; ICCR, International Cooperation on Cosmetics Regulation; iTTC, Internal Threshold of
 53 Toxicological Concern; KE, Key Event; KER, Key Event Relationship; log₁₀RBA, Logarithm (base 10)
 54 of Relative Binding Affinity; MIE, Molecular Initiating Event; MoA, Mode of Action; NGRA,
 55 Next-Generation Risk Assessment; NMPA, National Medical Products Administration; OECD,
 56 Organisation for Economic Co-operation and Development; PBPK, Physiologically Based
 57 Pharmacokinetic; PF, Problem Formulation; PR, Progesterone Receptor; QSAR, (Quantitative)
 58 Structure–Activity Relationship; REACH, Registration, Evaluation, Authorisation and Restriction of
 59 Chemicals; SCCP, Scientific Committee on Consumer Products; SCCS, Scientific Committee on
 60 Consumer Safety; SCHEER, Scientific Committee on Health, Environmental and Emerging Risks; TG,
 61 Test Guideline; TTC, Threshold of Toxicological Concern; WoE, Weight of Evidence.

62 **Introduction**

63 Since the European Union introduced its phased ban on animal testing for cosmetics under Directive
64 2003/15/EC, culminating in the full marketing ban under Regulation (EC) No 1223/2009 (EU 2003,
65 2009), next-generation risk assessment (NGRA) has become the institutional direction for cosmetic
66 safety assessment. The NGRA concept originated from the "toxicity testing in the 21st century" vision
67 (Krewski et al. 2010) and was established as the methodological framework for cosmetic ingredient risk
68 assessment through the nine ICCR principles (Dent et al. 2018). ICCR explicitly defined NGRA as an
69 "exposure-led, hypothesis-driven" paradigm integrating in silico, in chemico, and in vitro evidence, with
70 principles requiring exposure-driven assessment, hypothesis (mode of action, MoA/AOP)-oriented
71 strategies, and logically transparent documentation. The OECD guidance on integrated approaches to
72 testing and assessment (IATA) provided a framework for integrating heterogeneous evidence on the basis
73 of AOPs (OECD 2017), and Test Guideline (TG) 497 delivered the first example of a defined approach
74 that integrates in chemico, in vitro, and limited in silico information by fixed rules, gaining regulatory
75 acceptance for the skin sensitization endpoint (OECD 2021a). The SCCS Notes of Guidance further
76 incorporated NGRA logic and in silico assessment into ingredient safety evaluation (SCCS 2023).
77 However, while establishing these principles, ICCR also acknowledged that NGRA still "lacks
78 regulatory guidance on how to integrate heterogeneous data to support safety decisions" (Dent et al.
79 2018). Existing frameworks thus provide method access and basic principles but have not defined the
80 functional division and coordination logic of computational tools within the evidence chain, and have not
81 been translated into an executable assessment architecture.

82 This architectural gap manifests in the in silico layer as four interrelated structural deficits. First, the
83 hazard identification deficit: computational tools such as QSAR priority ranking, molecular docking for

mechanistic hypothesis generation, and integrated models assisting evidence adjudication are mostly deployed as atomized, parallel calls serving in vitro studies, without a systematic in silico hazard identification workflow or standardized applicability domain verification. Second, the exposure routing deficit: although ICCR listed exposure-led as a core NGRA principle, existing NGRA case studies mostly fix exposure information as a single-value parameter at the problem formulation stage, where it mainly performs entrance tiering and exit judgment functions such as defining exposure scenarios and verifying intake; it never becomes a routing variable that drives assessment-object reconstruction, so "exposure-led" remains confined to the entrance and exit of the assessment. Third, the AOP mapping deficit: the functional boundary of the in silico layer within the evidence chain is undefined, the correspondence between computational tools and molecular initiating events (MIEs)/AOPs is missing, and in silico evidence is treated as an adjunct to in vitro methods rather than as an evidence layer with a defined decision-making function. Fourth, the evidence integration deficit: cross-scale evidence lacks structured, auditable integration rules; even the regulatory-accepted defined approach of TG 497 integrates only limited in silico sources with in chemico and in vitro data for the single skin sensitization endpoint (OECD 2021a). These deficits do not stem from insufficient tool precision but from the absence of assessment architecture logic.

To address these deficits, this study reconstructs exposure information from an assessment parameter into a routing variable, replacing parallel tool calls with conditional dependence. In this design, the in silico layer acquires a defined upstream–downstream position relative to in vitro methods within the NGRA evidence chain: it completes hazard priority ranking and mechanism hypothesis generation before in vitro testing; when the evidence chain meets the credible negative threshold and is corroborated by external legacy evidence, it supports a credible negative conclusion without additional testing; and it

routes to targeted in vitro verification when evidence gaps are identified. On this basis, we construct a "screening–exposure–mechanism–decision" four-stage computational toxicology assessment paradigm, with progressive strategies and applicability domain verification targeting the hazard identification deficit, exposure-driven process routing targeting the exposure routing deficit, MIE-nature-driven three-mode functional mapping targeting the AOP mapping deficit, and WoE priority and conflict-resolution rules targeting the evidence integration deficit. The operability of the paradigm is demonstrated with homosalate (HMS) as a case study. The paradigm offers a computation-driven assessment pathway from hazard identification to decision output, providing a methodological reference for advancing NGRA from a principle-based framework toward operational implementation.

Capability Boundaries and Structural Deficits of Current In Silico Practice

Hazard Identification Deficit

NGRA practice employs quantitative structure–activity relationship (QSAR) models and structural alerts for hazard identification, but the capability boundaries of screening tools constitute a systematic identification blind spot. QSAR models are constrained by their applicability domain; when a target compound falls outside the training-set chemical space, prediction reliability decreases markedly. The VEGA platform incorporates an applicability domain index (ADI) that provides quantitative warnings (VEGA Hub 2024), and the OECD QSAR Toolbox dynamically constructs chemical categories through read-across (Dimitrov et al. 2016) to compensate for out-of-domain predictions; nevertheless, the applicability domain criteria of individual models remain independent and lack cross-model consistency (OECD 2024b). Structural alert tools cover only known mechanistic fragments: Cronin et al. (2022) showed that the predictive reliability of structural alerts varies with the strength of mechanistic evidence

and that they lack identification capability for novel structures or non-classical metabolic activation pathways. Screening tools also preset the assessment object as the parent compound, so metabolites are not covered; the precision of in silico metabolism prediction tools is only 1.1–29% across the six tools evaluated (Boyce et al. 2022), and metabolism-driven toxicity signals are systematically omitted at the screening stage. In practice, these limitations translate into imbalanced resource allocation and screening depth: applicability domain verification, multi-model integration (Collins et al. 2024), and progressive invocation by information depth have not become standardized operational requirements.

Exposure Routing Deficit

Existing NGRA practice has achieved exposure-driven tiered assessment, but exposure information primarily serves as an entrance parameter and an exit judgment. Berggren et al. (2017) proposed the SEURAT-1 three-tier workflow in which exposure serves as a Tier 0 entrance parameter for subsequent internal dose calculation; Alexander-White et al. (2022) introduced the threshold of toxicological concern (TTC) as a tiered exit point; and in the case studies on coumarin (Baltazar et al. 2020) and caffeine (Bury et al. 2021), exposure was fixed at the problem formulation stage. To transcend this limitation, this study defines "exposure routing" as the reconstruction of exposure information from a static parameter into a process-control variable that drives assessment pathway triage and dynamic expansion of the assessment object set; its signal forms comprise three classes. Systemic exposure potential, that is, the confirmation of dermal permeation and systemic exposure, determines whether systemic toxicity endpoints enter mechanistic assessment; exposure-entity identity and metabolic transformation, namely the fate of the parent and the identification of its metabolites, determine the composition of the assessment object set; and the dose–target relationship, the quantitative association between internal dose and target-site effects, determines the triage of assessment priority and subject

selection. The three signal classes are instantiated progressively according to the maturity of the available exposure data.

Existing practice has touched on individual elements of exposure routing without meeting its full requirements. Ebmeyer et al. (2024) identified the metabolite salicylic acid as the toxicological driver of benzyl salicylate, but metabolite inclusion was a forward sequential step rather than a mandatory loop back to the screening stage, and the switching of assessment objects still relied on ad hoc expert judgment rather than explicit rule-based computational triage. The EFSA TKPlate 1.0 platform selects different toxicokinetic modules on the basis of exposure data types (Dorne et al. 2023), but this routing is limited to computational precision selection within the same endpoint and does not involve cross-endpoint exposure routing or assessment-object switching. Teeguarden et al. (2016) proposed the aggregate exposure pathway (AEP) framework, which describes the path of exposure from source to target and intersects with AOPs at the MIE; however, AEP is a descriptive framework rather than normative process-control logic. Middleton et al. (2022) evaluated the protectiveness of non-animal systemic safety assessment in a toolbox-and-workflow form, with exposure information serving as input to the systemic safety judgment without driving assessment-object reconstruction or pathway triage. Finally, the ICCR principles require exposure characterization of impurities and metabolites (Dent et al. 2018), but this principle has not been translated into an executable operational pathway.

AOP Mapping Deficit

The AOP framework provides biological anchors for integrated assessment (OECD 2017). TG 497 validated at the skin sensitization endpoint that, when the AOP key event (KE) chain is complete and quantitative criteria are clear, the integration logic of multi-source predictions can be traced and verified

(OECD 2021a). Wittwehr et al. (2017) noted that AOP knowledge level affects the complexity of computational model selection but did not formalize the MIE knowledge state into functional triage logic. Villeneuve et al. (2014) proposed five core principles for AOP development and distinguished three stages of AOP maturity (putative, qualitative, quantitative). Verhoeven et al. (2024) proposed a WoE scoring method based on the Bradford–Hill criteria for quantitative confidence assessment of key event relationships (KERs) in AOP networks. The MechoA+ system classifies compounds by MIE mechanism and provides anchors for downstream QSAR modeling, but assigns compounds with unknown MIEs to a default narcosis class (Levet et al. 2026).

Differences in AOP maturity across endpoints imply that computational strategies should differ fundamentally. For skin sensitization, where the MIE is known and predictable, for phototoxicity, where the AOP has a hierarchical structure, and for eye irritation, where the MIE is a biologically consensual hypothesis of physicochemical damage, computational methods should respectively perform MIE identification and attribution, hierarchical screening, and downstream KE verification, rather than uniformly applying a fixed chain. None of the above work has formalized MIE nature into cross-endpoint functional triage logic.

At the IATA implementation level, existing work has likewise not established MIE-nature-driven mapping logic. The in silico toxicological problem formulation (PF) framework of Achar et al. (2024) reveals that methodological literature generally lacks high-level components such as conceptual models and problem formulation, with the biological basis and applicability boundaries of predictions often left implicit; among implemented IATAs, the skin sensitization defined approach is the only truly end-to-end, fixed-rule-driven implementation, while systemic toxicity-related IATA cases still rely mainly on expert judgment. Notably, even the GeoTox framework, which connects spatiotemporal exposure

characterization to individual- and population-level risk prediction, applies a uniform concentration–response and hazard quotient (HQ) mapping pipeline to different mechanistic endpoints in its implementation, so that mechanistic nature itself does not enter the mapping-strategy decision (Messier et al. 2025). The distinction lies between endpoints rather than within them: GeoTox applies one pipeline to all endpoints, whereas this paradigm triages endpoints by MIE nature into three modes with different operational logic and evidence weights (mechanistic attribution, hierarchical screening, statistical correlation inference).

Evidence Integration Deficit

Hazard identification, exposure assessment, and mechanistic analysis produce evidence of different types and confidence levels, and standardized rules for evidence weight assignment and conflict resolution are lacking. Existing integration practice relies mostly on expert judgment rather than on a traceable methodological framework. SCHEER (2024) and the EFSA Scientific Committee (2018) require transparency in evidence integration without providing operational rules: the EFSA uncertainty guidance does not differentiate the inherent reliability of computational evidence types, and the SCHEER WoE memorandum distinguishes evidence weight only qualitatively (e.g., classifying QSAR as supporting information), leaving no criteria for weighing QSAR predictions against molecular docking evidence. TG 497 validated the feasibility of rule-based evidence integration at the skin sensitization endpoint, but this defined approach combines OECD-validated in chemico and in vitro test data by fixed rules, is partially supplemented by limited in silico information in certain defined approaches, and is limited to a single endpoint (OECD 2021a).

Standard fragmentation further reinforces this deficit. At the regulatory level, REACH sets strict

conditions for QSAR acceptance under Annex XI §1.3, based on the OECD (Q)SAR validation five principles (OECD 2007), and ECHA guidance recommends using prediction consistency from multiple independent models to enhance acceptance confidence (ECHA 2025); the US EPA ToxCast program, by contrast, adopts a high-throughput screening prioritization strategy (Kavlock et al. 2012), with fundamental differences in acceptance logic and evidence standards between the two jurisdictions. Furthermore, SCCS and CIR can reach different safety conclusions for the same cosmetic ingredient because of different methodological choices and evidence acceptance standards. The benzophenone series exemplifies this: CIR, based on existing animal toxicological data, considered benzophenone-1, -2, and others safe (CIR 2021), whereas SCCS, based on new in vitro genotoxicity studies and endocrine disruption evidence, judged benzophenone-1 unsafe and could not reach a conclusion on benzophenone-2 (SCCS 2025a, b). At the methodological level, physiologically based pharmacokinetic (PBPK) verification standards have not reached consensus (OECD 2021b), the Cramer classification controversy and the internal TTC (iTTC) framework are still evolving (EFSA and WHO 2016; Patlewicz et al. 2022), and the BER exit threshold lacks a unified conclusion (Middleton et al. 2022). Methodological bottlenecks limit technical capability, while institutional barriers constrain regulatory acceptance willingness; the two reinforce each other.

The Exposure-Route-Driven Four-Stage Assessment Paradigm

Based on NGRA and IATA principles, this study constructs a "screening–exposure–mechanism–decision" four-stage computational assessment paradigm that organizes the originally isolated assessment scales of hazard identification, exposure assessment, mechanistic analysis, and decision integration into a progressive process. The paradigm follows four principles: risk-oriented, with assessment depth matched to risk intensity; mechanism-anchored, with computational predictions

anchored to the MIE and KEs of the AOP framework; evidence-progressive, with four-stage evidence advancing progressively; and transparent and traceable, with the assessment process fully recording evidence sources, verification status, and uncertainty boundaries. The methodological logic does not depend on specific software implementations; any qualifying tools of the same type can be substituted according to their applicability domains.

Stage 1 completes hazard identification and priority ranking, compensating for the hazard identification deficit through progressive strategies, applicability domain verification, and multi-model integration. Stages 2–4 respectively address the exposure routing, AOP mapping, and evidence integration deficits: exposure information is elevated from an assessment parameter to a process routing variable that drives dynamic assessment-object reconstruction through metabolite loop screening; a MIE-nature-driven functional positioning framework, moderated by AOP maturity with respect to the range of advance and evidence weight, is established; and WoE priority and conflict-resolution rules are defined. The four stages in series reconstruct the inter-tool relationship from parallel calls to conditional dependence, in which the output of the preceding stage determines the assessment objects and computational strategy of the subsequent stage; this constitutes the methodological core of the paradigm. The paradigm is benchmarked against the nine ICCR best practice principles (Dent et al. 2018), with the exposure-led principle that ICCR lists as the core of NGRA implemented through Stage 2 exposure routing, and the overall architecture achieves precise allocation of assessment resources through the pathway triage mechanism driven by exposure routing (Fig. 1).

Stage 1: Screening and Hazard Identification

Stage 1 is the entrance to the assessment chain and rapidly identifies substances requiring priority attention. Existing computational toxicology tools embody two prediction logics: structural alert logic, in

which specific structural fragments lead through known toxicity mechanisms to risk flags (qualitative prediction), and QSAR logic, in which overall molecular descriptors lead through statistical models to toxicity predictions (quantitative or semi-quantitative prediction). This paradigm does not stratify by tool type but by output information depth, adopting a progressive strategy of qualitative hazard profiling, quantitative prediction, and endpoint-specific screening.

Qualitative profiling determines whether risk signals exist. If a specific endpoint shows no positive signal, it does not enter quantitative prediction, conserving computational resources. Within the same positive endpoint, hazard intensity across compounds can vary by several orders of magnitude, requiring quantitative prediction to further differentiate priority. High-concern endpoints additionally require endpoint-specific screening tools for targeted expansion, covering mechanistic pathways not addressed by profiling and quantitative prediction. The predictive reliability of structural alerts is differentiated; Cronin et al. (2022) proposed grading the reliability of structural alerts for endpoints such as skin sensitization according to clearly defined criteria. The necessity of multi-model integration is empirically supported: voting methods demonstrated superiority in consensus modeling across multiple toxicological endpoints, and weighting methods are equally validated integration strategies (Collins et al. 2024); at the regulatory level, the OECD (Q)SAR assessment framework lists voting and weighting as accepted methods for integrating multi-model predictions (OECD 2024b). The paradigm accordingly recommends voting and weighting for integration. The paradigm further establishes a same-source deduplication rule: when the output of one tool occupies multiple evidence positions within the same stage (e.g., both as a standalone structural-alert prediction and as a model-panel member), its evidential weight is counted only once; the remaining positions are retained for reporting completeness but excluded from model consensus, avoiding double counting.

The screening stage calls computational tools progressively as needed. First, a combination of tools with different profiling logics is called for broad-spectrum hazard profiling, each providing structural alert, rule-oriented, or multi-source integration profiles that complement each other; if multiple tools produce inconsistent outputs for the same endpoint, the conservative principle flags it as uncertain; if profiling tools lack data coverage for an endpoint, supplementary calls to multi-model platforms such as VEGA are made. On the basis of profiling, quantitative or semi-quantitative prediction is initiated: the VEGA platform (incorporating the ADI applicability domain index; VEGA Hub 2024) and the OECD QSAR Toolbox (based on category-approach read-across and trend analysis; Dimitrov et al. 2016) can be used; when the target compound falls outside the VEGA model applicability domain, the Toolbox can dynamically construct categories from profiling results to fill the gap. High-concern endpoints further call endpoint-specific screening tools such as EDC Profiler for targeted expanded screening.

The above process produces hazard grading (positive/negative/uncertain) for each toxicological endpoint, MIE nature clues, and prediction confidence scores (Fig. 2). Uncertainty is judged from four dimensions (applicability domain coverage, training set representativeness, mechanistic consistency, and metabolic consideration) and rated low, medium, or high. The final grading of each endpoint, combined with product use patterns, determines the subsequent pathway. When any endpoint is uncertain or positive, it is flagged as high priority and proceeds to the next stage: Stage 2 for systemic toxicity endpoints or directly to Stage 3 for local toxicity endpoints (Fig. 2). When all endpoints are negative, or negative with low uncertainty, the compound is flagged as exposure-pending-assessment; for leave-on products or high-exposure scenarios it enters Stage 2 to investigate systemic toxicity mechanisms beyond screening-tool coverage, including metabolic activation; otherwise it is flagged as low priority and exits the assessment.

The triage pathway for high-priority endpoints depends on the spatial domain of the AOP to which they belong. For local toxicity endpoints (e.g., skin sensitization, skin irritation, eye irritation), the AOP cascade is anchored primarily in the microenvironment of the contact site, and its triggering does not depend on systemic absorption; therefore, when such endpoints are positive or uncertain, they are routed to Stage 3 for mechanistic assessment without waiting for dermal permeation evaluation results. Conversely, for systemic toxicity endpoints (e.g., endocrine disruption, organ toxicity, reproductive and developmental toxicity), the AOP cascade is anchored in internal target organs, and MIE triggering requires the chemical to enter the systemic circulation; therefore, when such endpoints are positive or uncertain, the dermal permeation evaluation result from Stage 2 is a prerequisite, and entry into Stage 3 occurs only after confirming significant systemic exposure.

Stage 2: Exposure Assessment and Access Determination

Stage 2 implements the triage direction from Stage 1 through dermal permeation assessment, metabolic prediction, and PBPK modeling (Fig. 2). Systemic toxicity endpoints that are positive or uncertain require confirmation of dermal permeation behavior: compounds that permeate enter Stage 3 for systemic toxicity AOP assessment, and those that do not permeate exit that endpoint. Routing for local toxicity endpoints does not depend on permeation results and follows the Stage 1 triage. For substances with all-negative screening but flagged as exposure-pending-assessment (leave-on products or high-exposure scenarios), dermal permeation prediction determines the next step: those that permeate are investigated for systemic toxicity mechanisms, and those that do not permeate exit.

Metabolite prediction identifies metabolites and loops back to Stage 1 for re-screening; the loop screening results and PBPK modeling jointly determine the assessment objects and mechanistic focus of Stage 3. Triage is dynamic: dermal permeation and metabolic behavior both vary with scenario

conditions such as product use patterns, and the same compound may enter different pathways in leave-on versus rinse-off products.

Dermal permeation assessment calls skin permeation models (e.g., SwissADME (Daina et al. 2017) and EPI Suite™ DERMWIN™ (US EPA 2012), both based on the Potts–Guy equation (Potts and Guy 1992)) to predict log K_p, and judges dermal permeation significance in combination with product use patterns. Inter-model prediction consistency is limited, and none of the models adequately accounts for evaporation effects (Grégoire et al. 2021), warranting combined multi-model use with cautious interpretation. To resolve such discrepancies, the paradigm applies an input-discrepancy disposition rule: experimental values take precedence; absent experimental data, source verification determines reconcilability; if irreconcilable, the conservative value is adopted. The K_p threshold is affected by inter-model differences, and the SCCS default 10% absorption rate (SCCS 2023) can serve as a conservative reference. Metabolic assessment sequentially calls CYP450 substrate probability prediction tools (e.g., ADMETlab (Fu et al. 2024), pkCSM (Pires et al. 2015)) and metabolite prediction tools (e.g., Meteor (Boyce et al. 2022), BioTransformer (Djoumbou-Feunang et al. 2019)) to identify metabolites and compare the systemic exposure levels of parent and metabolites. The overall reliability of metabolic prediction tools is limited (Boyce et al. 2022), with insufficient coverage of unconventional metabolic pathways; however, for metabolic pathways with clear reaction mechanisms (e.g., ester bond hydrolysis), the reliability of joint multi-tool prediction has been validated (Ebmeyer et al. 2024).

After metabolite identification, a loop back to Stage 1 re-executes hazard screening to investigate hazard signals not covered by the parent compound. The loop screening results determine the assessment objects and PBPK trigger conditions for Stage 3. Loop depth is by default limited to first-generation metabolites, because their systemic exposure contribution is typically significantly higher than that of

subsequent-generation metabolites; higher-generation metabolites are additionally screened only when PBPK modeling indicates that their exposure contribution is non-negligible. If the hazard profile of metabolites is consistent with that of the parent, PBPK modeling quantifies the systemic exposure difference between parent and metabolites to determine assessment priority. Whether the difference is sufficient to switch the assessment subject depends on endpoint-specific safety margins and cannot be settled by a single fold-threshold for all endpoints; when the difference is uncertain, the conservative default retains both parent and metabolites as assessment objects. Where metabolites exhibit hazard signals not present in the parent (e.g., a new target positive), the assessment must cover the new target and include the metabolites; PBPK then no longer has a decisive role in triage. This stage outputs pathway triage results, exposure subject lists, and metabolite loop screening results, providing the triage basis and assessment objects for Stage 3.

Stage 3: AOP-Oriented Mechanism-Driven In-Depth Assessment

In existing practice, computational methods uniformly apply fixed chains regardless of AOP maturity; prior work has addressed only the one-way association of AOP maturity or MIE nature without integrating the two or defining the correspondence between computational tools and MIE/AOP. The OECD IATA framework (OECD 2017) and AOP key event chain methodology (Wittwehr et al. 2017; Villeneuve et al. 2014) respectively provide the organizational form for test evidence and the structured description of AOPs, but neither provides rules for selecting computational methods by the epistemological state of the MIE. This study accordingly proposes a normative principle: when the MIE is known, computational methods should anchor to the MIE for mechanistic attribution; when it is unknown, they can only degenerate into statistical correlation inference. Stage 3 integrates these two aspects into a MIE-nature-driven functional positioning framework: MIE nature determines the

functional type of computational methods, and AOP maturity does not change functional attribution but moderates the range of advance and evidence weight of mechanistic assessment, transforming method selection from expert discretion into MIE-nature-driven conditional routing. Here, the MIE must be distinguished at two levels: endpoint-level MIE knowledge refers to the endpoint's mechanism-class affiliation documented in the literature and serves as the basis for mode assignment and verification design, whereas compound-level MIE triggering refers to the judgment of whether a specific compound triggers the MIE and is the execution target within a mode.

Endpoints entering this stage already carry MIE identification results progressively produced by the preceding stages. Qualitative profiling and quantitative prediction provide MIE nature clues for the endpoint (e.g., electrophilic structural alerts, receptor binding positive signals), and endpoint-specific screening further refines MIE identification depth. Under this design, MIE nature yields three functional modes (Fig. 3), whose assignment follows an explicit rule: an endpoint is assigned to Mode 1 when the MIE is known and computationally predictable (i.e., a structural or mechanistic model can reproduce the MIE trigger); to Mode 2 when the AOP exhibits a pre-MIE-to-MIE hierarchy that must be screened sequentially; and to Mode 3 otherwise (MIE unclear or AOP not formalized), with endpoints possessing a biologically consensual MIE retaining partial anchoring to the AOP chain. Mode 1 has regulatory precedents at both of its representative endpoints: at the skin sensitization endpoint, TG 497 combines OECD-validated in chemico and in vitro test data by fixed rules (OECD 2021a); at the endocrine disruption endpoint, the OECD has organized the available test methods into a tiered conceptual framework (OECD 2018). Mode 2 and Mode 3 are based on theoretical derivation from the AOP framework.

When the AOP MIE is known and computationally predictable (Mode 1), the function of computational methods is MIE identification and attribution, that is, determining whether the MIE is triggered and attributing positive results to specific mechanisms, as exemplified by the skin sensitization MIE of covalent binding between electrophilic chemicals and skin proteins, and the endocrine disruption MIE of ligand–receptor binding. When the AOP has a pre-MIE to MIE hierarchical structure (Mode 2), the function is hierarchical screening; for example, phototoxicity requires assessment of photoexcitation potential before identifying the type of photochemical reaction. When the MIE remains unclear, or only a biologically consensual hypothesis insufficient to anchor a computable AOP chain is available (Mode 3), the function degenerates into statistical correlation inference: computational methods cannot anchor to specific MIEs and can only provide indirect supporting evidence through network toxicology pathway enrichment and molecular docking. For endpoints with a biologically consensual MIE hypothesis (e.g., corneal epithelial cell toxicity in eye irritation), downstream key events can be selected for verification on the basis of that consensus, but the evidence weight is correspondingly downgraded and requires supplementary verification through the IATA framework before final judgment. Endpoints with a known MIE but no computationally predictable tools likewise execute as Mode 3, and the downgrade basis must be annotated in the mechanistic assessment report.

Mode attribution is determined by MIE nature; AOP maturity does not change mode attribution but consistently moderates evidence weight and range of advance across all three modes. Referring to the three-stage AOP maturity classification of Villeneuve et al. (2014), the paradigm operationalizes this moderation into two-level rules. In the range-of-advance dimension: a quantitative AOP (complete KE chain with dose–response and temporal consistency established) allows mechanistic assessment to proceed fully along the KE chain; a qualitative AOP (KE associations supported by evidence but lacking

quantitative relationships) limits the range of advance to the MIE and its directly adjacent downstream KEs, with anything beyond flagged as speculative; a putative AOP (KE associations based on hypothetical links) limits the range of advance to MIE identification and attribution, without supporting downstream KE verification. In the evidence-weight dimension: under a quantitative AOP, each mode retains its within-mode default weight; under a qualitative AOP, evidence weight is downgraded by one level (high→medium, medium→low); under a putative AOP, evidence weight is capped at low confidence. When the AOP knowledge base for a specific endpoint has atypical structures (e.g., jump-wise KE absence or weak KER evidence), the mechanistic assessment report must annotate the deviation and explain the adjustment basis, but the adjustment direction must be consistent with the above rules.

MIE identification and attribution may span multiple stages. When screening tools have already provided MIE-level identification results, Stage 3 completes mechanistic attribution from a structural perspective; when screening tools have already completed both identification and attribution, Stage 3 has no advanceable analytical target for that endpoint, and the mechanistic assessment report only records the completion status and basis of identification and attribution, with that endpoint proceeding directly to the Stage 4 WoE integration.

Representative endpoints of Mode 1 include skin sensitization (covalent binding MIE, with structural alerts directly identifying and attributing) and endocrine disruption (receptor binding MIE, with dual-platform target prediction for intersection, pathway enrichment, and molecular docking progressing sequentially, each layer executed depending on the evidence output of the preceding stages, and molecular docking as the core layer). Upon positive docking, the dynamic stability of the complex can be further assessed through molecular dynamics simulation (van der Spoel et al. 2005; Case et al. 2025).

Mode 2 is represented by phototoxicity, where the pre-MIE is photoexcitation potential and the MIE is the type of photochemical reaction; this hierarchical structure has been formally expressed in the OECD No. 397 IATA framework (OECD 2024a), providing a regulatory precedent for the hierarchical screening function of Mode 2. Mode 3 encompasses two categories of endpoints: those with unclear MIEs (e.g., chronic systemic toxicity, reproductive and developmental toxicity) and those with biologically consensual MIE hypotheses but lacking formalized AOPs (e.g., eye irritation). For the former, computational methods rely purely on statistical correlation inference. For the latter, the consensual MIE provides partial biological anchors, allowing the selection of downstream key events for targeted verification, thereby offering a larger operational space than pure statistical inference. However, both categories lack formal AOP chain constraints and carry correspondingly downgraded evidence weights.

Across all three modes, both systemic and local toxicity endpoints can be distributed according to their MIE nature, and mechanistic assessment methods are tailored accordingly. Endpoints with a known MIE but no computationally predictable tools likewise execute as Mode 3, and the downgrade basis must be annotated in the mechanistic assessment report. This stage outputs a mechanistic assessment report, providing quantifiable and traceable mechanistic evidence for WoE integration.

Stage 4: WoE and IATA Decision Integration

Stage 4 establishes evidence weight priority and conflict-resolution rules, using WoE methodology as the evidence integration logic and the IATA framework (OECD 2017, 2024a) as the assessment strategy organization (Fig. 4).

As a design choice of the paradigm rather than an empirically validated weighting scheme, the default WoE weight priority, from high to low, is: Stage 3 mechanistic verification evidence; Stage 2 exposure

parameters, PBPK modeling results, and metabolite loop screening signals; Stage 1 in-domain multi-model consensus quantitative QSAR predictions and multi-tool consistent positive signals from endpoint-specific screening; out-of-domain or single-tool low-confidence predictions. This ordering rests on three design rationales: mechanistic anchoring, as higher-level evidence is associated with identified MIEs or verified exposure pathways; exposure constraint, as exposure-layer signals constrain the biological relevance of hazard predictions; and multi-model robustness, as multi-model consensus outperforms single-tool outputs. The levels of the ladder serve different functions: mechanistic verification evidence directly supports hazard-direction adjudication, whereas exposure parameters and metabolite loop screening signals modulate the assessment objects and priority and constrain the biological relevance of hazard predictions without independently adjudicating hazard direction; the default priority of the ladder applies to evidence weighting and concern configuration, while cross-scale directional conflicts are handled by the conflict-resolution rules. The weight of qualitative structural alerts is assigned to the appropriate level on the basis of their confidence assessment results, rather than being listed separately as the lowest priority. The premise of this priority is that the confidence of upper-level evidence is no lower than that of lower-level evidence; when cross-scale evidence shows quality inversion, a weighting principle of confidence over hierarchy is adopted. In this rule, confidence is graded within each evidence type (Table 1), whereas the hierarchy reflects typical cross-type reliability; under inversion, the within-type confidence rather than the type's typical rank governs.

The four-dimensional uncertainty assessment defined in Stage 1 addresses the quality of individual tool predictions; the confidence assessment in Stage 4 integrates the verification depth of cross-scale evidence on top of the uncertainty assessment, through a graded mapping of evidence type and uncertainty level to three-level confidence. The cross-evidence-type confidence assessment uses a

qualitative three-level framework (Table 1). Although confidence across different evidence types cannot be numerically normalized, evidence within the same level is comparable, providing an adjudication anchor for conflict resolution. The specific values of quantitative thresholds depend on tool parameters and the chemical space of the specific endpoint and therefore require case-specific calibration; the decision rules rather than the parameter values constitute the fixed component of the paradigm.

Conflict-resolution rules distinguish directional conflicts from intensity conflicts. Directional conflicts refer to different-scale evidence pointing to opposite conclusions, in which high-confidence evidence at the mechanistic level is prioritized. Intensity conflicts refer to consistent directions but significant differences in evidence confidence or resolution, in which the evidence with both confidence and resolution advantages serves as the main body, with deviations recorded in the uncertainty boundary. If both types of conflict cannot be resolved, IATA directs targeted supplementary verification. After rule-based conflict resolution, if residual uncertainty does not meet the threshold for the final conclusion, historical experimental data from regulatory databases can be introduced for independent verification; such data must satisfy three conditions: no new animal testing; use only to confirm a computational tendency as credible negative, never to flip negative to positive; and sourcing from public assessment documents accepted by regulatory authorities. This asymmetric arrangement is anchored in the mechanistic asymmetry between the two directions of evidence: negative evidence is mechanistically homogeneous and highly specific, so that once established it supports an exclusion judgment, whereas positive evidence is mechanistically heterogeneous with diverse false-positive sources, so that endpoint establishment requires convergence of multiple evidence lines and mechanistic anchoring. When verification results are consistent with the computational tendency, the determination is confirmed; when inconsistent, the case is escalated to IATA-directed supplementary verification. Within the same scale,

higher verification depth carries greater weight. The precedence rule for heterogeneous evidence is as follows: when in vitro experimental data and same-source in silico predictions coexist for the same endpoint, the experimental data take precedence, and the in silico predictions serve to explain discrepancies and indicate directions without overriding the experimental judgment.

The IATA framework organizes the evidence from the first three stages according to the logic of information needs, evidence supply, and decision nodes. Screening provides hazard identification evidence, exposure assessment provides exposure pathway and dose evidence, and mechanistic assessment provides biological plausibility evidence, forming a progressive evidence chain. When evidence is insufficient to support a decision, IATA guides the targeted triggering of subsequent testing strategies, specifying which endpoint and which scale requires supplementary data. After in vitro test results are returned to the overall NGRA/IATA framework, the AOP mechanistic anchoring and cross-endpoint evidence integration provided by this paradigm offer the mechanistic structure for organizing in vitro data and incorporating them into the overall evidence weight. If systemic effect signals are discovered along the KE chain during AOP analysis of local toxicity endpoints, they are flagged as items of concern in the WoE, and IATA triggers targeted systemic toxicity assessment. Ultimately, three levels of assessment conclusions are formed: Credible Negative, where evidence from all scales (screening–exposure–mechanism) consistently indicates low risk, key evidence confidence reaches medium–high level, and conflict resolution confirms that no unresolved directional or intensity conflicts remain; Positive with Concern, where some evidence suggests potential risk requiring IATA-guided targeted supplementation; and Clear Positive, where cross-scale evidence consistently indicates high risk (a paradigm-internal decision category assigned on computational evidence, pending

external confirmation). The entire assessment process records evidence sources, verification status, and uncertainty boundaries.

Paradigm Validation: Four-Stage Assessment of Homosalate

Homosalate (HMS, CAS 118-56-9) is an organic UV filter commonly used in leave-on sunscreen products. The following case study demonstrates the operability of the four-stage paradigm constructed above.

In Silico Methods and Parameters

The case study was executed with the following tools (full protocols in the Supplementary Information). Structural alert screening was performed with Toxtree v3.1.0; multi-endpoint hazard profiling with the US EPA Hazard Cheminformatics Module (HCM); quantitative endpoint predictions with the VEGA platform v1.2.6 (six skin sensitization models) and US EPA WebTEST; and nuclear receptor screening with the EDC Profiler web server. Dermal permeation was predicted with EPI Suite™ DERMWIN™ and SwissADME. CYP substrate specificity was assessed with ADMETlab 3.0 and Deep-PK, and Phase I oxidative metabolites were generated with BioTransformer. Molecular docking was performed with AutoDock 4/MGL Tools 1.5.7 (Lamarckian genetic algorithm; grid box $40 \times 40 \times 40$ Å, spacing 0.375 Å) using the ERα ligand-binding domain (PDB 3ERT) and the PR ligand-binding domain (PDB 1A28). Tool outputs and prediction tables are provided in Tables S1–S4 and Figs. S1–S9.

Multi-Endpoint Hazard Profiling and Priority Determination

Hazard screening employed a multi-tool progressive strategy. Toxtree structural alerts excluded the skin sensitization MIE (negative) and detected a phenolic structure suggesting potential endocrine disruption

(Fig. S1), but structural alerts provide only qualitative judgment and cannot quantify endpoint intensity, and their predictive reliability must be explicitly assessed against established criteria (Cronin et al. 2022). The US EPA Hazard Cheminformatics Module (HCM) multi-endpoint hazard profiling added the quantitative dimension, rating endocrine disruption and developmental toxicity as high concern, although the platform did not cover skin sensitization, irritation, and eye irritation, which are core endpoints for cosmetic safety (Fig. S2; Vegosen and Martin 2020). For the key gap of skin sensitization, six skin sensitization models from the VEGA platform (five QSAR models and one rule-based system) were called for prediction; statistical and mechanistic models showed directional conflict and the endpoint was flagged as uncertain (Table S1). Note that Toxtree contributes to both the standalone structural-alert screen and the VEGA model panel; to avoid double counting of the same evidence, its output is counted only once, namely as the standalone structural-alert (mechanistic) prediction, with the panel entry retained for reporting completeness but excluded from the model consensus. The two occurrences of this output (standalone structural alert and panel entry) are both negative and stem from the same prediction; the uncertainty arises from the directional conflict between this mechanistic negative and the positive predictions of the statistical VEGA models (CAESAR, NCSTOX). For the endocrine disruption endpoint, already identified as high concern by HCM, further refinement to specific receptors was needed: WebTEST (US EPA 2025) and EDC Profiler (Qiu et al. 2026) directionally confirmed that HMS was positive on four nuclear receptors, namely androgen receptor (AR, antagonist), estrogen receptor α (ER α , agonist), estrogen receptor β (ER β , antagonist), and glucocorticoid receptor (GR, antagonist) (Figs. S3-S4). The endocrine disruption effect, as the primary endpoint, entered Stage 2, while the skin sensitization uncertainty was resolved at the WoE stage (no advanceable mechanistic target; Sect. 4.4).

Dermal Permeation Confirmation and Oxidative Metabolite Loop Screening

The preceding hazard screening confirmed that HMS possesses endocrine disruption potential against multiple nuclear receptors; whether this molecular hazard translates into in vivo health risk largely depends on its systemic exposure level. In the dermal permeation assessment, DERMWIN™ (EPI Suite) (US EPA 2012) and SwissADME (Daina et al. 2017) both predicted the skin permeation coefficient (log K_p) of HMS on the basis of the Potts–Guy equation (Potts and Guy 1992). Owing to the equation's high sensitivity to lipophilicity input parameters, the two results differed substantially: DERMWIN™, based on its estimated log K_{ow} (6.16), predicted K_p = 0.657 cm/h (log K_p = –0.18), whereas SwissADME, based on its underlying XLOGP3 value (5.05), predicted K_p = 0.178 cm/h (log K_p = –0.75), an approximately 3.7-fold difference. Applying the input-discrepancy disposition rule, HMS lacks an experimentally determined log K_{ow}, and source verification confirmed that the discrepancy originates from differing lipophilicity inputs and cannot be reconciled; the lower permeation coefficient (log K_p = –0.75, SwissADME) was therefore adopted as the conservative value. Notably, even the conservative lower prediction (log K_p = –0.75) places HMS near the upper end of the experimental log K_p range (–6.11 to –0.19) of the 106-compound dataset compiled from the Flynn and Wilschut databases (Chen et al. 2018). Given that HMS is widely used as a UV absorber in leave-on sunscreen cosmetics, its high predicted dermal permeability implies that HMS readily enters the systemic circulation, satisfying the exposure premise for systemic-toxicity-related AOPs to enter mechanistic assessment.

Once in the systemic circulation, the metabolic behavior of HMS becomes crucial. ADMETlab 3.0 (Fu et al. 2024) and Deep-PK (Myung et al. 2024) consistently predicted HMS as a CYP2C9 substrate with high confidence (Fig. S5). Based on this consensus, BioTransformer (Djoumbou-Feunang et al. 2019) was called to simulate CYP-mediated Phase I oxidative metabolism. The predictions showed that HMS

can generate ten first-generation mono-oxidation metabolites, classifiable by oxidation site into three categories: aromatic ring hydroxylation, alicyclic ring hydroxylation, and methyl hydroxylation; the metabolites are designated C1–C10 (Table S2).

Beyond the oxidative pathway, GLORYx (de Bruyn Kops et al. 2021) was used to predict Phase I metabolic reactions, identifying an ester bond hydrolysis pathway that yields salicylic acid and 3,3,5-trimethylcyclohexanol (Table S3); this prediction is consistent with the hydrolysis behavior of homosalate documented in the SCCS/1638/21 scientific advice (SCCS 2021b). According to the assessment-object reconstruction rule of the paradigm, the two hydrolysis products were included in the assessment list. External ToxCast data (Onyango et al. 2023) showed that salicylic acid was active in the ATG_AR_TRANS_up reporter gene assay (AC50 1.003 μ M, ACC 1.093 μ M, far below the cytotoxicity threshold of 1000 μ M), whereas 3,3,5-trimethylcyclohexanol was inactive. Both hydrolysis products already have ToxCast/Tox21 high-throughput screening experimental data, whose evidential status is higher than that of same-source in silico predictions; this study therefore did not perform separate in silico screening for the hydrolysis products, and the evidence of this branch relies primarily on external experimental data. The AR transactivation signal of salicylic acid is consistent with the conclusion of Ebmeyer et al. (2024) based on the zebrafish model that the salicylic acid structural unit is the key driver of the endocrine-disrupting activity of benzoate esters; the decision-category assignment of this branch is given in the Stage 4 WoE integration.

In the cosmetic NGRA context, high-quality PBPK parameters for HMS and its metabolites along the skin–systemic circulation axis are insufficient, and precise quantification of the systemic exposure contribution of the parent and each metabolite involves significant uncertainty. According to the paradigm triage logic, when exposure differences are uncertain, a conservative default strategy retains

both parent and metabolites as assessment objects; accordingly, this study included all ten first-generation oxidative metabolites together with the parent compound in the assessment list, to avoid under-evaluation and to examine the reshaping effect of metabolic transformation on the hazard profile. For signal attribution, the dermal permeation confirmation was the exposure premise for the endocrine disruption endpoint to enter mechanistic assessment. The assessment-object reconstruction was jointly driven by the routing signals of metabolic transformation and metabolite identification, together with the hazard signal differences identified by loop screening. The dose–target relationship did not participate in this triage owing to the absence of PBPK; its routing function awaits PBPK pathway validation. EDC Profiler screening showed (Fig. 5) that the most significant change was the consistent shift of PR activity from inactive (parent) to active (10/10 metabolites), indicating that CYP-mediated oxidative metabolism significantly enhanced the PR interference signal relative to the parent compound. The high 10/10 consistency does not exclude the contribution of a shared systematic bias of the two tools, such as overrepresentation of agonist-type structures in the training chemical space; the signal is therefore interpreted as a predicted binding signal awaiting external verification rather than as confirmation of enhanced activity. The responses of the other nuclear receptors showed clear site dependence: aromatic ring hydroxylation products exhibited elevated ER α activity, whereas alicyclic and methyl hydroxylation products showed loss of GR activity (Figs. S6-S7). Independent prediction by WebTEST corroborated this structure–activity trend, confirming that the hydroxyl position determines the direction of ER affinity (aromatic ring enhancement vs. alicyclic ring attenuation; Fig. S8). The consistent activation of PR by metabolites and the site-dependent differentiation of ER affinity together indicate that the hazard profile based on the parent compound may not fully represent the toxicological characteristics of the

chemical entities actually exposed after metabolism. The hydrolysis-branch objects were calibrated with external ToxCast data, forming two mutually independent evidence lines with the oxidation branch.

ER α Binding Mode Differences and the PR Transition

The first two stages identified two endpoints requiring advancement at the mechanistic level. For skin sensitization, the uncertainty arose from directional conflict between VEGA statistical and mechanistic models, and MIE attribution (electrophilic structural alert) had already been determined as negative by Toxtree, leaving no advanceable target for mechanistic assessment. For endocrine disruption, the ligand–receptor binding MIE belongs to the Mode 1 receptor-binding subclass, requiring attribution of the initial positive result to a specific receptor. ER α was selected as the docking target because this endpoint has WebTEST quantitative prediction available for cross-referencing (Fig. S8); representative metabolites were the aromatic ring hydroxylation product (C6) with elevated ER affinity and the alicyclic hydroxylation product (C8) with decreased ER affinity, while the methyl hydroxylation product (C9) was not included because of activity similar to the parent.

AutoDock 4 (Morris et al. 2009) was used to dock HMS, C6, and C8 with ER α (PDB 3ERT; Shiau et al. 1998), yielding a binding affinity ranking of C6 (–5.79 kcal/mol) > HMS (–5.63 kcal/mol) > C8 (–5.39 kcal/mol), consistent with the WebTEST prediction trend ($\log_{10}$ RBA: –1.38 > –2.06 > –2.94). The binding mode differences focused on the anchoring of key residues Glu353 and Arg394 (Fig. S9): the para-hydroxyl of the C6 aromatic ring, directed toward Glu353, formed a double anchor that stabilized the ligand–receptor complex, whereas the C8 alicyclic hydroxyl deviated from the polar region, maintaining only the single-end binding of the parent.

Loop screening revealed that PR is the only target with a qualitative difference between metabolites and parent (parent negative; all ten oxidative metabolites positive), requiring validation at the receptor binding level. Molecular docking of HMS, C6, C8, and C9 with the PR ligand-binding domain (PDB 1A28; Williams and Sigler 1998) yielded a binding affinity ranking of C6 (−6.53 kcal/mol) > C8 (−6.18 kcal/mol) > HMS (−6.04 kcal/mol) > C9 (−5.79 kcal/mol) (Table S4). In the binding modes, the C6 aromatic ring para-phenolic hydroxyl formed double hydrogen bonds with GLN725 (2.1 Å) and LEU718 (1.9 Å), being the main contributor to its strongest binding; the C8 alicyclic hydroxyl, although forming three hydrogen bonds with GLY722, LEU718, and ASN719, had its binding weakened by the long GLY722 bond distance; and the C9 methyl hydroxyl formed two hydrogen bonds with LEU718 and ASN719, with the weakest binding. All four compounds, including the parent predicted as negative, could form stable interactions with the PR binding cavity; molecular docking thereby attributed the metabolite PR interference signal to the ligand–receptor binding MIE, but the stable binding of the parent indicates that the docking evidence cannot discriminate between the presence and absence of activity between the parent and metabolites, providing structural support rather than activity verification.

Three-Endpoint WoE Integration and Conflict Resolution

The evidence from the first three stages was integrated by endpoint according to the WoE rules. For the endocrine disruption endpoint, the ER α evidence chain was the most complete: initial screening by WebTEST and EDC Profiler was directionally consistent positive (high confidence); loop screening metabolite activity changes were consistent with the QSAR trend (medium confidence); and molecular docking validated the structural basis (high confidence). With confidence levels within each evidence type assigned according to Table 1, the three-stage evidence was directionally consistent and progressively deepened, and was judged Clear Positive according to the three-level conclusion definition.

Within the ER α evidence chain, EDC Profiler provided qualitative classification probability while WebTEST provided quantitative affinity prediction (log₁₀RBA); the two were consistent in reflecting the ER α binding strength trend, but WebTEST was superior to the EDC Profiler classification threshold in quantitative resolution, constituting an intensity conflict based on evidence resolution. According to the intensity conflict-resolution rule, the high-confidence WebTEST quantitative trend served as the main body, with the insufficient EDC Profiler resolution recorded in the uncertainty boundary. AR, ER β , and GR did not enter mechanistic assessment verification and retained only screening-level evidence, flagged as items of concern.

For the skin sensitization endpoint, the Toxtree mechanistic negative (high confidence) and the VEGA statistical model positive (medium–low confidence) constituted a directional conflict; according to the directional conflict-resolution rule, the high-confidence mechanistic level was prioritized, and the computational framework tended toward negative. The high-confidence assignment of Toxtree is bounded by the mechanistic coverage of its rule set: its skin sensitization rules are anchored in the MIE of covalent binding between electrophilic substances and skin proteins, so the negative result can only exclude sensitization potential mediated by structural alerts covered by the rule library, not sensitization risk mediated by other mechanisms outside its coverage. However, the credible negative determination requires evidence from all scales to consistently indicate low risk, and the internal directional conflict of computational evidence prevented the consistency threshold from being met, requiring external evidence to break the impasse. The historical animal experimental data in the SCCP opinion on homosalate (no sensitization observed in existing animal data; SCCP 2007) served as independent verification and was directionally consistent with the computational tendency, confirming the tendency toward negative as credible negative. This historical data constitutes legacy evidence in the regulatory database and serves

only the external calibration function for the computational paradigm: the computational framework, driven by in silico evidence, completed the resolution from uncertain to tending negative, and the legacy evidence confirmed the tending negative as credible negative without new animal testing, reflecting the evidence screening and triage role of the paradigm within the IATA framework.

For the metabolite PR interference, molecular docking provided structural support rather than activity verification; the Positive with Concern judgment was based on the loop screening signal, with IATA-guided targeted in vitro progesterone receptor activity verification recommended. The identification of the metabolite PR prediction signal is an in silico output of exposure-routing-driven assessment-object reconstruction, and the relative intensity between the metabolite and the parent cannot be determined from the screening output. For the hydrolysis branch, the AR transactivation signal of salicylic acid (external ToxCast data, with activity far below the cytotoxicity threshold) was flagged as Positive with Concern. With PBPK absent, its systemic exposure implication cannot be quantified, and the in vitro activity cannot yet be converted into an in vivo risk judgment; its exposure contribution awaits PBPK-path quantification. This treatment is consistent with the claim that metabolites must be considered, and the hydrolysis products are retained at the conclusion level. Integrating all endpoints, the WoE conclusions for HMS comprise four items covering all three output pathways of the paradigm: ER α was judged Clear Positive; skin sensitization was judged Credible Negative; and both the metabolite PR interference and the AR signal of salicylic acid were judged Positive with Concern, with the former requiring targeted supplementary in vitro verification and the latter awaiting exposure quantification. These conclusions demonstrate the integration logic of the computational paradigm as the in silico layer of the NGRA evidence chain (Fig. 6). They are all paradigm-internal decision categories assigned on the

basis of computational evidence; external (in vitro or regulatory) confirmation is required before any regulatory conclusion is drawn.

External Calibration: Case-Critical Predictions against Published Experimental Data

External calibration I: the prediction–experiment discrepancy of the PR signal. Schreurs et al. (2005) reported in reporter gene assays that HMS acts as an antagonist at both the androgen receptor (AR) and the progesterone receptor (PR), with most effects occurring above 1 μ M. The experimental antagonism is not contradictory to the stable binding found by molecular docking: binding is a necessary but not sufficient condition for agonist-mode activity. The discrepancy may originate from a training chemical space of the prediction tools biased toward agonists or from failure to distinguish agonist and antagonist modes; this hypothesis has yet to be verified against the tool documentation. The loop screening results should be interpreted as a predicted binding signal that the metabolites retain PR activity: because the screening output failed to reflect the antagonistic activity of the parent, the relative intensity between metabolites and parent exceeds the resolution of the present screening, being neither a biological transition of presence/absence nor an intensity ranking supported by the available evidence. The Positive with Concern judgment is retained, and the in vitro verification directed by IATA should be designed to discriminate between agonist and antagonist modes.

External calibration II: comparison of the ER α judgment with ToxCast/Tox21 and human plasma data. Reanalysis of 453 ToxCast/Tox21 assays for HMS showed activity in 42 assays (9.3%), mostly above the cytotoxicity burst threshold, with the most sensitive ER-related assay (ATG_ERE_CIS_up) at AC50 14.20 μ M and ACC 8.86 μ M (Onyango et al. 2023). Compared with the human plasma C_{max} of 0.029 μ M after a single application and 0.088 μ M under maximum-use conditions, the safety margin is approximately 100–490-fold. The authors of that study, affiliated with the safety assessment department

of a sunscreen manufacturer, interpreted the activity above the burst threshold as more likely reflecting general cytotoxicity. This study separates the data from the interpretation: the AC50/ACC values and the plasma Cmax data are adopted, whereas the authors' interpretive judgment of activity attribution is not, and the same data are independently interpreted here. This calibration does not contradict the SCCS position that the endocrine disruption evidence for HMS remains inconclusive (SCCS 2021a); rather, it exemplifies the type of external evidence against which the paradigm-internal ER α category assigned in this study must ultimately be tested, rather than a confirmation of that category.

Challenges and Perspectives

The remaining boundaries and breakthrough pathways are discussed below from four dimensions: verification depth, tool reliability, AOP knowledge coverage, and regulatory acceptance.

Uneven Verification Depth

The HMS case instantiated two of the three routing signal classes, namely systemic exposure potential and exposure-entity identity with metabolic transformation: metabolite loop screening triggered assessment-object reconstruction and revealed a PR interference signal predicted significantly stronger in the metabolites than in the parent, while the routing function of the dose–target relationship was not instantiated owing to the absence of PBPK. Among the three-mode mappings, Mode 1 is empirically validated. Mode 2 has a regulatory precedent in the OECD No. 397 IATA framework for phototoxicity, which endorses a tiered testing strategy; however, the specific functional mapping proposed here, in which computational tools are sequentially routed along the pre-MIE to MIE hierarchy, has not yet been empirically validated. Mode 3 remains purely theoretical derivation. The functional logic of Modes 2 and 3 can be indirectly supported at endpoints where the AOP knowledge base already has entries: the

pre-MIE to MIE hierarchical structure of the phototoxicity AOP supports the feasibility of the Mode 2 hierarchical screening logic, and the biological consensus MIE of eye irritation supports the feasibility of Mode 3's consensus-based downstream KE verification; nevertheless, complete functional mapping validation still requires execution within a unified framework. Paths not covered by the case validation include the non-permeation pathway, the all-negative screening pathway, and the triage logic that quantifies exposure differences through PBPK when metabolite hazard profiles are consistent with the parent. The paradigm has currently been validated for only a single compound, and the applicability of exposure routing and three-mode mapping in mixture scenarios has not been assessed; a single case can inherently cover only one triage pathway. The breakthrough pathway proceeds in two steps. In the near term, supplementary PBPK pathway validation should be prioritized: SCCS-assessed ingredients whose metabolite hazard profiles are consistent with the parent can be re-evaluated through the four-stage paradigm, and the consistency rate of the PBPK triage logic with the original conclusions quantified. In the medium-to-long term, a standardized test set should be constructed covering metabolite–parent differences, endpoints with different AOP maturity, screening–mechanism assessment conflicts, endpoints with incomplete AOPs, and mixture scenarios, evolving toward formulation-level assessment and drawing on the OECD TG 497 validation paradigm.

Tool Reliability Ceiling at Routing Key Nodes

The triage logic and three-mode mapping of the paradigm both take the prediction signals of computational tools as input; the validity of the architectural logic is independent of tool precision, whereas the reliability of the triage judgments it produces is constrained by signal quality. Inter-model prediction consistency for dermal permeation is limited (Grégoire et al. 2021), the precision of metabolite prediction tools is only 1.1–29% (Sect. 2.1), and the K_p threshold and metabolite exposure

contribution triage thresholds have not been robustly assessed across chemical spaces. The immediately actionable improvement is to implement multi-model ensembles and mandatory applicability domain verification at routing key nodes: triage judgments should be confirmed by at least two independent model consistencies, with conservative default thresholds for Kp and metabolite exposure contribution (SCCS 2023); simultaneously, sensitivity analysis on these thresholds should be conducted to determine the robust interval of triage conclusions under parameter perturbation. The medium-to-long-term direction is to establish dedicated validation benchmarks for routing key nodes, quantify prediction error distributions, and move triage thresholds from empirical judgment toward statistical calibration. Furthermore, the uncertainty of a single computational module propagates nonlinearly through the four-stage chain; developing probabilistic uncertainty quantification methods that convert deterministic outputs into confidence interval distributions is a direction for further improving paradigm reliability, with the current qualitative grading rules providing a scalable foundational framework.

AOP Knowledge Coverage and Standardization of Fallback Strategies

The three-mode mapping presupposes that MIE nature can be determined. Endpoints such as skin sensitization and phototoxicity have known MIEs and formalized AOPs, falling into Modes 1 and 2; endpoints such as eye irritation have a biological consensus MIE but lack a formalized AOP, and endpoints such as chronic systemic toxicity and reproductive and developmental toxicity have MIEs that remain unclear, both falling into Mode 3 (statistical correlation inference), where the mechanistic anchoring principle partially fails. For some endpoints where the MIE is known but the KE chain is incomplete, Mode 1 can still be entered for MIE identification and attribution, but the range of advance is limited and the evidence weight correspondingly downgraded. Mode 3 has been incorporated into the mapping system, but the judgment criteria for its supporting evidence still require cross-endpoint case

799 accumulation and calibration. The breakthrough pathway has two layers. For Mode 3 judgment rule
800 calibration, cases of statistical correlation inference across different endpoints need to be accumulated to
801 quantify the sensitivity and robustness of threshold parameters; for AOP knowledge base expansion,
802 reliance on the continuous collaborative development of the AOP community is needed to upgrade more
803 endpoints from Mode 3 to Mode 1 or Mode 2.

804 **Regulatory Acceptance Distance**

805 The mechanistic evidence chain of Stage 3 and the WoE/IATA integration conclusions of Stage 4 are
806 methodologically self-consistent, but they have no explicit mapping to regulatory acceptance standards.
807 REACH sets five validation conditions for QSAR predictions (OECD 2007), and SCCS requires in vivo
808 data or equivalent alternative evidence (SCCS 2023); the SCCS evaluation history of the endocrine
809 disruption properties of homosalate was formed according to this logic (SCCS 2021a, b). Its 2021
810 opinion considered the available evidence on endocrine disruption "inconclusive, and at best equivocal"
811 and narrowed the maximum concentration from 10% to 0.5% (SCCS 2021a); in the same year, the
812 scientific advice for facial products, after supplementing with PBPK modeling, allowed 7.34% (SCCS
813 2021b). Whether the computational prediction level of this paradigm can directly satisfy the above
814 acceptance conditions remains undetermined. The pathway to shortening the acceptance distance is joint
815 research with regulatory authorities, mapping the evidence output of each stage of the paradigm to SCCS
816 and REACH acceptance conditions and forming a three-level acceptance mapping table of "directly
817 acceptable/requires in vitro supplementation/requires in vivo supplementation"; the paradigm provides a
818 decomposable and traceable evidence structure as the basis for such joint research.

819 **Conclusion**

This study reconstructed exposure information from an assessment parameter into a routing variable and formalized a "screening–exposure–mechanism–decision" four-stage computational toxicology assessment paradigm, in which MIE-nature-driven three-mode functional mapping defines the biological anchors of computational methods and WoE priority and conflict-resolution rules enable transparent integration of heterogeneous evidence, positioning the in silico layer as a structured evidence contributor within the NGRA evidence chain. In the homosalate case, exposure routing expanded the assessment object from the parent to its metabolites, loop screening identified a progesterone receptor interference prediction signal in the oxidative metabolites, demonstrating the core mechanism of exposure-routing-driven assessment-object reconstruction; both directional and intensity conflict-resolution rules were effectively triggered. The current boundaries are that only Mode 1 of the three-mode mapping is empirically validated while Modes 2 and 3 remain theoretical derivation, and that the regulatory acceptance distance has not yet been bridged, awaiting standardized test sets and joint research with regulators. The paradigm provides a computation-driven assessment pathway from hazard identification to decision output for cosmetic safety assessment and a decomposable, traceable methodological path for advancing NGRA from a principle-based framework toward operational implementation; the generality of the paradigm-internal decision categories awaits external confirmation.

Author contributions Conceptualization: Jihui Fang, Qunyue Wu; Methodology: Fangfang Zhou, Jingwen Liang, Dong Guo, Yangjie Li; Writing – original draft: Fangfang Zhou, Dong Guo; Writing – review & editing: Dong Guo, Fangfang Zhou, Yanjia Cai; Funding acquisition: Dong Guo, Jihui Fang, Jingwen Liang.

Funding Open Access funding provided thanks to the Scientific and technological innovation project

of Guangdong Provincial Drug Administration (2025YDZ13, 2026JDZ04) and Guangdong Provincial Medical Science and Technology Research Fund (B2026149).

Data availability All data generated or analyzed during this study are included in this published article and its supplementary information files.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Reference

- Achar J, Cronin MTD, Firman JW et al (2024) A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment. *Arch Toxicol* 98:1727–1740. <https://doi.org/10.1007/s00204-024-03721-6>
- Alexander-White C, Bury D, Cronin M et al (2022) A 10-step framework for use of read-across (RAX) in next generation risk assessment (NGRA) for cosmetics safety assessment. *Regul Toxicol Pharmacol* 129:105094. <https://doi.org/10.1016/j.yrtph.2021.105094>
- Baltazar MT, Cable S, Carmichael PL et al (2020) A next-generation risk assessment case study for coumarin in cosmetic products. *Toxicol Sci* 176(1):236–252. <https://doi.org/10.1093/toxsci/kfaa048>
- Berggren E, White A, Ouedraogo G et al (2017) Ab initio chemical safety assessment: a workflow based on exposure considerations and non-animal methods. *Comput Toxicol* 4:31–44. <https://doi.org/10.1016/j.comtox.2017.10.001>

863 Boyce M, Meyer B, Grulke C et al (2022) Comparing the performance and coverage of selected in silico
 864 (liver) metabolism tools relative to reported studies in the literature to inform analogue selection in
 865 read-across: a case study. *Comput Toxicol* 21:100208.
 866 <https://doi.org/10.1016/j.comtox.2021.100208>
 867 Bury D, Alexander-White C, Clewell HJ 3rd et al (2021) New framework for a non-animal approach
 868 adequately assures the safety of cosmetic ingredients—a case study on caffeine. *Regul Toxicol*
 869 *Pharmacol* 123:104931. <https://doi.org/10.1016/j.yrtph.2021.104931>
 870 Case DA, Cerutti DS, Cruzeiro VWD et al (2025) Recent developments in AMBER biomolecular
 871 simulations. *J Chem Inf Model* 65(15):7835–7843. <https://doi.org/10.1021/acs.jcim.5c01063>
 872 Chen C-P, Chen C-C, Huang C-W et al (2018) Evaluating molecular properties involved in transport of
 873 small molecules in stratum corneum: a quantitative structure-activity relationship for skin
 874 permeability. *Molecules* 23(4):911. <https://doi.org/10.3390/molecules23040911>
 875 CIR Expert Panel (2021) Amended safety assessment of benzophenones as used in cosmetics. Final
 876 amended report, 6 April 2021. *Cosmetic Ingredient Review*, Washington, DC.
 877 <https://www.cir-safety.org/sites/default/files/benzop032021FAR.pdf>
 878 Collins SP, Mailloux B, Kulkarni S et al (2024) Development and application of consensus in silico
 879 models for advancing high-throughput toxicological predictions. *Front Pharmacol* 15:1307905.
 880 <https://doi.org/10.3389/fphar.2024.1307905>
 881 Cronin MTD, Bauer FJ, Bonnell M et al (2022) A scheme to evaluate structural alerts to predict toxicity—
 882 assessing confidence by characterising uncertainties. *Regul Toxicol Pharmacol* 135:105249.
 883 <https://doi.org/10.1016/j.yrtph.2022.105249>

884 Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics,
885 drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep* 7(1):42717.
886 <https://doi.org/10.1038/srep42717>

887 de Bruyn Kops C, Šícho M, Mazzolari A et al (2021) GLORYx: prediction of the metabolites resulting
888 from phase 1 and phase 2 biotransformations of xenobiotics. *Chem Res Toxicol* 34(2):286–299.
889 <https://doi.org/10.1021/acs.chemrestox.0c00224>

890 Dent M, Amaral RT, Da Silva PA et al (2018) Principles underpinning the use of new methodologies in
891 the risk assessment of cosmetic ingredients. *Comput Toxicol* 7:20–26.
892 <https://doi.org/10.1016/j.comtox.2018.06.001>

893 Dimitrov SD, Diderich R, Sobanski T et al (2016) QSAR Toolbox–workflow and major functionalities.
894 SAR QSAR Environ Res 27(3):203–219. <https://doi.org/10.1080/1062936X.2015.1136680>

895 Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A et al (2019) BioTransformer: a comprehensive
896 computational tool for small molecule metabolism prediction and metabolite identification. *J*
897 *Cheminform* 11(1):2. <https://doi.org/10.1186/s13321-018-0324-5>

898 Dorne JLCM, Cortiñas-Abrahantes J, Spyropoulos F et al (2023) TKPlate 1.0: an open-access platform
899 for toxicokinetic and toxicodynamic modelling of chemicals to implement new approach
900 methodologies in chemical risk assessment. *EFSA J* 21(11):e211101.
901 <https://doi.org/10.2903/j.efsa.2023.e211101>

902 Ebmeyer J, Najjar A, Lange D et al (2024) Next generation risk assessment: an ab initio case study to
903 assess the systemic safety of the cosmetic ingredient, benzyl salicylate, after dermal exposure. *Front*
904 *Pharmacol* 15:1345992. <https://doi.org/10.3389/fphar.2024.1345992>

905 ECHA (2025) Practical guide: how to use and report (Q)SARs, version 3.1.1. European Chemicals
 906 Agency, Helsinki. <https://doi.org/10.2823/8315768>

907 EFSA and WHO (2016) Review of the threshold of toxicological concern (TTC) approach and
 908 development of new TTC decision tree. EFSA Supporting Publication 13(3):EN-1006.
 909 <https://doi.org/10.2903/sp.efsa.2016.EN-1006>

910 EFSA Scientific Committee (2018) Guidance on uncertainty analysis in scientific assessments. EFSA J
 911 16(1):5123. <https://doi.org/10.2903/j.efsa.2018.5123>

912 EU (2003) Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003
 913 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States
 914 relating to cosmetic products. Off J Eur Union L66:26–35

915 EU (2009) Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30
 916 November 2009 on cosmetic products. Off J Eur Union L342:59–209

917 Fu L, Liu L, Li S et al (2024) ADMETlab 3.0: an updated comprehensive online ADMET prediction
 918 platform enhanced with broader coverage, improved performance, API functionality and decision
 919 support. Nucleic Acids Res 52(W1):W422–W431. <https://doi.org/10.1093/nar/gkac236>

920 Grégoire S, Sorrell I, Lange D et al (2021) Cosmetics Europe evaluation of 6 in silico skin penetration
 921 models. Comput Toxicol 19:100177. <https://doi.org/10.1016/j.comtox.2021.100177>

922 Kavlock R, Chandler K, Houck K et al (2012) Update on EPA's ToxCast program: providing high
 923 throughput decision support tools for chemical risk management. Chem Res Toxicol 25(7):1287–
 924 1302. <https://doi.org/10.1021/tx3000939>

925 Krewski D, Acosta D Jr, Andersen M et al (2010) Toxicity testing in the 21st century: a vision and a
 926 strategy. *J Toxicol Environ Health B* 13(2–4):51–138.
 927 <https://doi.org/10.1080/10937404.2010.483176>
 928 Levet G, Bauer FJ, Thomas PC et al (2026) MechoA+: a chemical structure profiler raising the bar for the
 929 prediction of mechanisms of toxic action for chemical safety assessment. *Environ Sci Technol*
 930 60(24):17124–17135. <https://doi.org/10.1021/acs.est.5c18657>
 931 Messier KP, Reif DM, Marvel SW (2025) The GeoTox package: open-source software for connecting
 932 spatiotemporal exposure to individual and population-level risk. *Hum Genomics* 19(1):5.
 933 <https://doi.org/10.1186/s40246-024-00711-8>
 934 Middleton AM, Reynolds J, Cable S et al (2022) Are non-animal systemic safety assessments protective?
 935 A toolbox and workflow. *Toxicol Sci* 189(1):124–147. <https://doi.org/10.1093/toxsci/kfac068>
 936 Morris GM, Huey R, Lindstrom W et al (2009) AutoDock4 and AutoDockTools4: automated docking
 937 with selective receptor flexibility. *J Comput Chem* 30(16):2785–2791.
 938 <https://doi.org/10.1002/jcc.21256>
 939 Myung Y, de Sá AGC, Ascher DB (2024) Deep-PK: deep learning for small molecule pharmacokinetic
 940 and toxicity prediction. *Nucleic Acids Res* 52(W1):W469–W475.
 941 <https://doi.org/10.1093/nar/gkae254>
 942 OECD (2007) Guidance document on the validation of (quantitative) structure-activity relationships
 943 [(Q)SAR] models. OECD Series on Testing and Assessment, No. 69. ENV/JM/MONO(2007)2.
 944 <https://doi.org/10.1787/9789264085442-en>

945 OECD (2017) Guidance document for the use of adverse outcome pathways in developing integrated
 946 approaches to testing and assessment (IATA). OECD Series on Testing and Assessment, No. 260.
 947 OECD Publishing, Paris. <https://doi.org/10.1787/44bb06c1-en>

948 OECD (2018) Revised guidance document 150 on standardised test guidelines for evaluating chemicals
 949 for endocrine disruption, 2nd edn. OECD Series on Testing and Assessment, No. 150. OECD
 950 Publishing, Paris. <https://doi.org/10.1787/9789264304741-en>

951 OECD (2021a) Test No. 497: defined approaches on skin sensitisation. OECD Guidelines for the Testing
 952 of Chemicals, Section 4. OECD Publishing, Paris. <https://doi.org/10.1787/b92879a4-en>

953 OECD (2021b) Guidance document on the characterisation, validation and reporting of physiologically
 954 based kinetic (PBK) models for regulatory purposes. OECD Series on Testing and Assessment, No.
 955 331. OECD Publishing, Paris. <https://doi.org/10.1787/d0de241f-en>

956 OECD (2024a) Guidance document on integrated approaches to testing and assessment (IATA) for
 957 phototoxicity testing. OECD Series on Testing and Assessment, No. 397. OECD Publishing, Paris.
 958 <https://doi.org/10.1787/8a979653-en>

959 OECD (2024b) (Q)SAR assessment framework: guidance for the regulatory assessment of (quantitative)
 960 structure activity relationship models and predictions, 2nd edn. OECD Series on Testing and
 961 Assessment, No. 405. OECD Publishing, Paris. <https://doi.org/10.1787/bbdac345-en>

962 Onyango DO, Selman BG, Rose JL et al (2023) Comparison between endocrine activity assessed using
 963 ToxCast/Tox21 database and human plasma concentration of sunscreen active ingredients/UV
 964 filters. Toxicol Sci 196(1):25–37. <https://doi.org/10.1093/toxsci/kfad082>

965 Patlewicz G, Worth A, Yang C et al (2022) Editorial: advances and refinements in the development and
 966 application of threshold of toxicological concern. *Front Toxicol* 4:882321.
 967 <https://doi.org/10.3389/ftox.2022.882321>
 968 Pires DEV, Blundell TL, Ascher DB (2015) pkCSM: predicting small-molecule pharmacokinetic and
 969 toxicity properties using graph-based signatures. *J Med Chem* 58(9):4066–4072.
 970 <https://doi.org/10.1021/acs.jmedchem.5b00104>
 971 Potts RO, Guy RH (1992) Predicting skin permeability. *Pharm Res* 9(5):663–669.
 972 <https://doi.org/10.1023/A:1015810312465>
 973 Qiu Y, Cui S, Zhang H et al (2026) EDC Profiler: a high-throughput screening server for identifying
 974 endocrine disrupting chemicals toward multiple nuclear receptors. *Environ Sci Technol*
 975 60(19):14023–14034. <https://doi.org/10.1021/acs.est.5c18135>
 976 SCCP (2007) Opinion on homosalate (COLIPA n° S12). SCCP/1086/07. Scientific Committee on
 977 Consumer Products.
 978 https://ec.europa.eu/health/archive/ph_risk/committees/04_sccp/docs/sccp_o_097.pdf
 979 SCCS (2021a) Opinion on homosalate (CAS No 118-56-9, EC No 204-260-8). SCCS/1622/20. Scientific
 980 Committee on Consumer Safety. https://health.ec.europa.eu/publications/homosalate_en
 981 SCCS (2021b) Scientific advice on the safety of homosalate (CAS No 118-56-9, EC No 204-260-8) as a
 982 UV-filter in cosmetic products. SCCS/1638/21. Scientific Committee on Consumer Safety.
 983 https://health.ec.europa.eu/system/files/2022-08/sccs_o_260.pdf
 984 SCCS (2023) SCCS notes of guidance for the testing of cosmetic ingredients and their safety evaluation,
 985 12th revision. SCCS/1647/22. Scientific Committee on Consumer Safety.

986 https://health.ec.europa.eu/publications/sccs-notes-guidance-testing-cosmetic-ingredients-and-their-safety-evaluation-12th-revision_en

987

988 SCCS (2025a) Opinion on benzophenone-1 (CAS No. 131-56-6, EC No. 205-029-4). SCCS/1672/24.

989 Scientific Committee on Consumer Safety, Brussels.

990 https://health.ec.europa.eu/publications/sccs-opinion-benzophenone-1-cas-no-131-56-6-ec-no-205-029-4_en

991

992 SCCS (2025b) Scientific advice on the safety of benzophenone-2 and benzophenone-5 as substances

993 with potential endocrine disrupting properties in cosmetic products. SCCS/1679/25. Scientific

994 Committee on Consumer Safety, Brussels.

995 https://health.ec.europa.eu/publications/sccs-scientific-advice-benzophenone-2-bp-2-and-benzophenone-5-bp-5-substances-potential-endocrine_en

996

997 SCHEER (2024) Memorandum on weight of evidence approach for risk assessment–revision 2024.

998 Scientific Committee on Health, Environmental and Emerging Risks, European Commission.

999 https://health.ec.europa.eu/publications/scheer-memorandum-weight-evidence-approach-risk-assessment-revision-2024_en

1000

1001 Schreurs RH, Sonneveld E, Jansen JH et al (2005) Interaction of polycyclic musks and UV filters with

1002 the estrogen receptor (ER), androgen receptor (AR), and progesterone receptor (PR) in reporter

1003 gene bioassays. Toxicol Sci 83(2):264–272. <https://doi.org/10.1093/toxsci/kfi035>

1004 Shiau AK, Barstad D, Loria PM et al (1998) The structural basis of estrogen receptor/coactivator

1005 recognition and the antagonism of this interaction by tamoxifen. Cell 95(7):927–937.

1006 [https://doi.org/10.1016/S0092-8674\(00\)81717-1](https://doi.org/10.1016/S0092-8674(00)81717-1)

1007 Teeguarden JG, Tan YM, Edwards SW et al (2016) Completing the link between exposure science and
 1008 toxicology for improved environmental health decisions: the aggregate exposure pathway
 1009 framework. *Environ Sci Technol* 50(9):4579–4586. <https://doi.org/10.1021/acs.est.5b05311>
 1010 US EPA (2012) Estimation Programs Interface Suite™ (EPI Suite™), version 4.11. US Environmental
 1011 Protection Agency, Washington, DC.
 1012 <https://www.epa.gov/tsca-screening-tools/epi-suite-estimation-program-interface>
 1013 US EPA (2025) User's guide for WebTEST (version 1.0) (web-services toxicity estimation software tool).
 1014 <https://comptox.epa.gov/dashboard/predictions>
 1015 van der Spoel D, Lindahl E, Hess B et al (2005) GROMACS: fast, flexible, and free. *J Comput Chem*
 1016 26(16):1701–1718. <https://doi.org/10.1002/jcc.20291>
 1017 VEGA Hub (2024) VEGA platform for QSAR modeling. <https://www.vegahub.eu/>
 1018 Vegosen L, Martin TM (2020) An automated framework for compiling and integrating chemical hazard
 1019 data. *Clean Technol Environ Policy* 22(2):441–458. <https://doi.org/10.1007/s10098-019-01795-w>
 1020 Verhoeven A, van Ertvelde J, Boeckmans J et al (2024) A quantitative weight-of-evidence method for
 1021 confidence assessment of adverse outcome pathway networks: a case study on chemical-induced
 1022 liver steatosis. *Toxicology* 505:153814. <https://doi.org/10.1016/j.tox.2024.153814>
 1023 Villeneuve DL, Crump D, Garcia-Reyero N et al (2014) Adverse outcome pathway (AOP) development
 1024 I: strategies and principles. *Toxicol Sci* 142(2):312–320. <https://doi.org/10.1093/toxsci/kfu199>
 1025 Williams SP, Sigler PB (1998) Atomic structure of progesterone complexed with its receptor. *Nature*
 1026 393(6683):392–396. <https://doi.org/10.1038/30775>
 1027 Wittwehr C, Aladjov H, Ankley G et al (2017) How adverse outcome pathways can aid the development
 1028 and use of computational prediction models for regulatory toxicology. *Toxicol Sci* 155(2):326–336.

1030 **Table of Contents and Figure captions**

1031 **Fig. 1** The exposure-route-driven four-stage in silico assessment paradigm within the NGRA/IATA
1032 framework. Stages 1–4 proceed sequentially from hazard landscape construction through exposure-route
1033 decision and MIE-driven mechanism assessment to confidence-weighted WoE integration, with
1034 exposure operating as the routing variable at Stage 2; their outputs are integrated into three decision
1035 categories (Clear Positive, Positive with Concern, Credible Negative).

1036 **Fig. 2** Exposure-route-driven pathway triage and metabolite-mediated reassessment logic (Stages 1–2).
1037 Stage 1 outputs are triaged by spatial-domain classification into three branches: local toxicity endpoints
1038 (positive or uncertain) enter Stage 3 directly through the local AOP pathway; systemic toxicity
1039 endpoints (positive or uncertain) pass a dermal penetration gate followed by a metabolism gate;
1040 exposure-dependent endpoints (all negative but high-exposure scenario) pass a dermal penetration gate
1041 followed by a systemic relevance judgement, and failing either gate leads to low priority. Metabolites
1042 identified in prediction trigger re-entry into Stage 1 for hazard reassessment and update of the
1043 assessment object set, so that the parent and relevant metabolites proceed to Stage 3 as local or
1044 systemic AOP pathways.

1045 **Fig. 3** MIE-nature-driven three-mode functional mapping of computational methods (Stage 3). Mode 1
1046 (MIE known and predictable) attributes MIE occurrence and yields direct mechanistic evidence; Mode
1047 2 (pre-MIE-to-MIE hierarchy within AOP) screens sequentially along the hierarchy and yields partial
1048 AOP pathway evidence; Mode 3 (MIE undefined or not formalized) infers statistical associations and
1049 yields supportive evidence. AOP maturity (quantitative, qualitative, putative) moderates the actionable
1050 range and evidence weight without altering mode assignment.

1051 **Fig. 4** Weight-of-evidence integration and conflict-resolution framework (Stage 4). Evidence is

weighted along the default priority ladder (mechanistic verification, exposure–PBPK–metabolite-loopback evidence, in-domain multi-model QSAR consensus and endpoint-specific screening, out-of-domain or low-confidence predictions), with confidence overriding hierarchy under cross-scale quality inversion; consistent evidence proceeds directly to the three decision categories, directional and intensity conflicts are resolved by their respective rules, and unresolved conflicts trigger IATA-directed supplementary testing.

Fig. 5 Nuclear receptor activity profiles of HMS (parent) and its ten first-generation oxidative metabolites predicted by EDC Profiler. PR is inactive in the parent but active in all ten metabolites (10/10), indicating that CYP-mediated oxidation endows HMS with progesterone-receptor interference absent in the parent.

Fig. 6 HMS case study mapped onto the four-stage paradigm as three representative endpoint pathways: ER α agonism (Clear Positive, high confidence), skin sensitization (Credible Negative, medium–high confidence; direction conflict resolved by high-confidence mechanistic evidence, with SCCP legacy support), and metabolite PR shift (Positive with Concern, moderate confidence; targeted in vitro PR validation under IATA).

Table 1. Cross-evidence-type confidence assessment criteria

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1.docx](#)
- [20260916ESI.docx](#)
- [Figures.pdf](#)