Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AOPedex: AOP-based endocrine disruption extrapolation as white-box mechanism maps for zero-data species

Yanpeng Gao

gaoyp2016@gdut.edu.cn

Guangdong University of Technology

Yanqin Zhang

Guangdong University of Technology

Guanhui Chen

Guangdong University of Technology

Na Luo

Guangdong University of Technology

Yuemeng Ji

Guangdong University of Technology

Yang-Guang Gu

South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, China

https://orcid.org/0000-0003-2314-0463

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11579523/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

AOPedex: AOP-based endocrine disruption extrapolation as white-box mechanism maps for zero-data species

Yanqin Zhang[1], Guanhui Chen[1], Na Luo[1], Yanpeng Gao[1*], Yuemeng Ji[1], Yang-Guang Gu[2*]

[1]Guangdong-Hong Kong-Macao Joint Laboratory for Contaminants Exposure and Health, Guangdong Key Laboratory of Environmental Catalysis and Health Risk Control, Institute of Environmental Health and Pollution control, Guangdong University of Technology, Guangzhou 510006, China

[2]South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510300, China

Correspondence: Yanpeng Gao (gaoyp2016@gdut.edu.cn) and Yangguang Gu (hydrobio@163.com).

Abstract

Endocrine disruption evidence is concentrated in a handful of model species, leaving risk assessment for 'zero-data species' unresolved; scarce toxicity records and biologically untraceable black-box extrapolation constrain machine learning. We present AOPedex, a knowledge-driven white-box framework embedding ToxCast in vitro activity, receptor binding-pocket conservation and traceable close-relative evidence in a deterministic AOP event-network propagation skeleton, so that mechanistic maps share one source for prediction, attribution and audit. Across 5,387 compounds and 756 vertebrate species, 94.4% of species are high-confidence or predictable, including 123 IUCN threatened species. Under identical protocols, EBM matches TabPFN in macro out-of-fold ROC-AUC on seven general-toxicity endpoints (0.810 versus 0.812) and outperforms it on data-sparse Morphology (0.698 versus 0.661). Controlled AON polarity randomisation barely changes ROC-AUC but disrupts mechanistic ordering, showing that predictive performance does not imply mechanistic validity. AOPedex audits receptor inputs, AON propagation, model readouts and literature evidence, yielding applicability-domain-conditioned mechanistic predictions, candidate screening regions and testable molecular entry points and event chains. Experimental feedback revises priors for iterative computation–experiment calibration. AOPedex aims to turn observations of model species into knowledge of many more so that toxicological knowledge serves biodiversity.

Introduction

Endocrine-disrupting chemicals (EDCs) interfere with natural hormone action and disrupt endocrine homeostasis in wildlife, causing abnormal gonadal development, reduced fecundity and, ultimately, population decline[1,2]. In a whole-lake experiment, environmentally relevant concentrations of 17α-ethynylestradiol (EE2) drove fathead minnow to reproductive collapse and its population to the brink of extirpation[3]; cascading through the food web, these losses depressed top predators, delayed population-level recovery and even reshaped reproductive social behaviour[4–6]. Population models predict that polychlorinated biphenyl (PCB) burdens will crash more than half of the world's killer whale populations within a century[7]; exposure to perfluorinated compounds already exceeds immunotoxicity thresholds in most polar bear subpopulations, and their long-chain homologues continue to accumulate in the Arctic[8]. For threatened species, endocrine-disrupting effects bear on survival itself.

Which threatened species are experiencing endocrine-disrupting effects has yet to be established at scale, and the evidence that does exist is extremely unevenly distributed across species. In the EPA ECOTOX Knowledgebase, 100 species contribute half of all records while more than 9,000 species are documented by only one or two publications each[9], with the records centred on standard laboratory organisms such as zebrafish, fathead minnow and rat. Amphibians and reptiles lack even mature test methods and assessment frameworks[10–12], and many threatened species have no endocrine-toxicity records at all. Filling these gaps with animal testing is unsustainable on ethical, cost and time grounds alike[13]. What is needed instead is computational toxicology that extrapolates toxicity from model to non-model organisms with precision, interpretability and broad coverage.

Machine learning offers new tools for this task[13], but it depends on these same unevenly sampled data. The systematic skew of existing toxicity databases makes records from different species and endpoints difficult to combine directly. Species and endpoint coverage in ECOTOX has blind spots[14], and the chemical space of ToxCast is biased towards EPA priority categories[15]. Selection bias introduced by data availability, experimental choice and the publication process separates observed associations from causal evidence[16,17], and black-box models trained on such data produce attribution that is vague and unauditable. For threatened species, a single misjudgement can misdirect conservation action, so any model used to inform decisions must withstand case-by-case scrutiny. Tiered frameworks for using available evidence have been proposed[18], but

they remain conceptual, with computational implementation and empirical testing still missing. The immediate task is therefore to implement these tiered evidence frameworks and test them in case-by-case assessments.

Faced with this data vacuum, a natural route is transfer learning, in which models are first trained on data-rich domains and then transferred to data-scarce target domains. Environmental toxicology recognised the route early[19], and recent work shows that it can succeed. Liang et al.[20] predicted the acute toxicity of 30 per- and polyfluoroalkyl substances (PFAS) to 181 aquatic species using the ICE database, and Wang et al.[21] applied generic chemical-structure representations to PFAS receptor-activation prediction. When target-domain data are extremely scarce, however, transferable representations are hard to calibrate against target-species labels, and the larger the cross-domain distance, the faster the transfer gains decay[22]. Meta-learning offers another path. Schlender et al.[23] used auxiliary data from other species to improve prediction for low-resource species, and ToxiSpecies[24] meta-trains across species so that models adapt to new tasks from a handful of samples. Under zero-shot conditions, though, where the target species carries no labels at all, inner-loop adaptation that relies on labelled support sets loses its supervised basis. The model can only extrapolate from its meta-trained priors, calibrating neither species-specific toxicity mechanisms nor its own features. Zero labels also mean that no target-domain validation can empirically corroborate the outputs; the plausibility of a prediction must then rest on scrutiny of its basis, and post hoc explanations alone do not establish the biological basis of extrapolation[25]. Transparency must instead be built into the structure of the model itself.

Mechanistic knowledge provides a basis for extrapolation when target-species labels are unavailable for calibration. Many endocrine processes are highly conserved across vertebrates. Aromatase inhibition, for example, depresses oestrogen levels and impairs reproduction[26]. Whether such conservation supports extrapolation to fish, birds, reptiles or mammals, however, varies by pathway and taxon, and depends on species-specific evidence at the level of receptor sequence and downstream events[10]. SeqAPASS grounds cross-species extrapolation at the target level in structural biology by comparing protein sequences and key-residue conservation[27], and growing genome and transcriptome resources steadily widen the range of comparable species[13]. Domain knowledge has likewise long been used to constrain machine learning. Informed machine learning has produced a taxonomy of strategies for injecting prior knowledge[28], physics-informed machine learning has shown that mechanistic constraints can sharply reduce data requirements[29], and the strategy has been codified for environmental applications[30]. These paradigms, however, have not been

systematically implemented for cross-species extrapolation in toxicology. We refer to this path as knowledge-driven modelling.

Adverse outcome pathway (AOP) networks are such a framework, encoding cross-species toxicological knowledge as causal chains of events. The concept was systematically articulated[31] and has since entered regulatory practice[32,33]. Haigis et al.[34] manually integrated sequence conservation with empirical evidence to delineate the taxonomic applicability domain of a thyroid-disruption AOP network, whose cross-species applicability domain still had to be established through expert review. Computational attempts are emerging. Xin et al.[35] embedded AOP priors in a knowledge graph to predict androgen-receptor-mediated reproductive toxicity in zebrafish, but the underlying knowledge curation relies on manual expert work, and transfer to distantly related taxa was not validated. Tan et al.[36] injected causal constraints into graph learning for multi-endpoint prediction, yet the modelling stays on the compound side and does not touch cross-species extrapolation. In all of these systems, however, knowledge curation and evidence registration remain manual expert work, a bottleneck that LLM agents are beginning to relieve by extracting structured evidence directly from the scientific literature[37]. Each of these efforts captures one link, be it receptor conservation, cross-species transfer or causal constraints, but none integrates all three within a unified AOP network framework with traceable, falsifiable mechanistic outputs.

To fill this gap, we present AOP-based Endocrine Disruption EXtrapolation (AOPedex), a knowledge-driven framework. Using the AOP mechanistic network (AON) as its backbone[38,39], AOPedex takes ToxCast in vitro activities[40] and species-specific receptor binding-pocket conservation[27,41] as inputs, and produces traceable mechanistic pathways and mechanistic maps of endocrine disruption. The causal event chain of an AOP is itself an interpretable structure, and as a prediction propagates through the event network, every step corresponds to a concrete event with its evidence. When a target species lacks its own receptor sequences, the model builds genus-level proxies or receptor-stratified k-NN inputs from sequenced close relatives, recording evidence provenance throughout. Under a unified chemical catalogue and outer split, it trains 13 task-specific white-box supervised heads based on explainable boosting machines. Each prediction decomposes into an exact sum of feature-term contributions, and mismatched samples can be audited layer by layer along the AOP event chain. Where environmental machine-learning models use post hoc explanations, AOPedex makes the white-box readout the inference itself. Attribution fidelity and predictive accuracy are computed by the same functions, so mechanistic node-level intervention and row-by-row auditing become native outputs of the

model. On classification endpoints, the white-box model matches TabPFN in discrimination (macro OOF ROC-AUC 0.810 versus 0.812) and slightly overtakes it on the data-sparse morphology endpoint. Applicability-domain assessment covers 756 vertebrate species, including 138 IUCN threatened species, of which 123 reach predictable or high-confidence states. For zero-data species, AOPedex provides predicted endocrine-disrupting effects as maps carrying mechanistic pathways, confidence and applicability boundaries.

Results

AOPedex takes compound activities from ToxCast, species receptor conservation from binding-pocket sequence alignments of 1,118 species, and signal-transduction paths from the AOP mechanistic network. The event network serves as the base layer of the mechanistic map, on which AOPedex builds auditable prediction heads. Under a single outer split, we evaluate the discrimination and calibration of five endocrine continuous endpoints, one thyroid binary endpoint and seven general-toxicity classification endpoints in parallel; ablation experiments test the contribution of each component, and we examine parameter stability under alternative reference-species parameterisations together with the taxonomic-distance dependence of mechanistic-vector reconstruction error. Applicability-domain analysis covers 756 vertebrate species. Ten case cards demonstrate the layer-by-layer auditability of the mechanistic maps, with the audit loop checking predictions, attributions and evidence conflicts and every revision verified through versioned recomputation.

The AOPedex framework, from mechanistic propagation to auditable prediction

We merged 173 endocrine-related AOPs from AOP-Wiki into a 629-node directed mechanistic network (AON; Figure 1). All 711 relationship edges are traceable to their sources; 541 are supported by adjacent key event relationships (KERs) and qualify for stepwise propagation, while non-adjacent summary relations are retained for display only (Methods). Nodes are organised into three levels — molecular initiating events (MIEs), key events (KEs) and adverse outcomes (AOs) — and support six independently supervised endocrine endpoints, five continuous potency endpoints (oestrogen receptor (ER) agonism, ER antagonism, androgen receptor (AR) agonism, AR antagonism and H295R steroidogenesis) and one binary endpoint (hTPO inhibition), together covering the oestrogenic, androgenic, thyroidal and steroidogenic (EATS) axes. Seven general-toxicity endpoints (three human KEs and four zebrafish-embryo AOs) provide parallel auxiliary supervision. The backbone is extended to 12 receptors. Species inputs come from the species' own pocket sequences or traceable close-relative proxies; receptors lacking a reliable MIE, such as the

progesterone receptor (PGR), do not enter the mechanistic entry layer. The species-side input is the binding-pocket residue conservation of ligand-binding domains for 756 species × 12 receptors, aggregated through a genus-level representative panel (751 genera) at the key-residue resolution used by SeqAPASS.

Figure 1

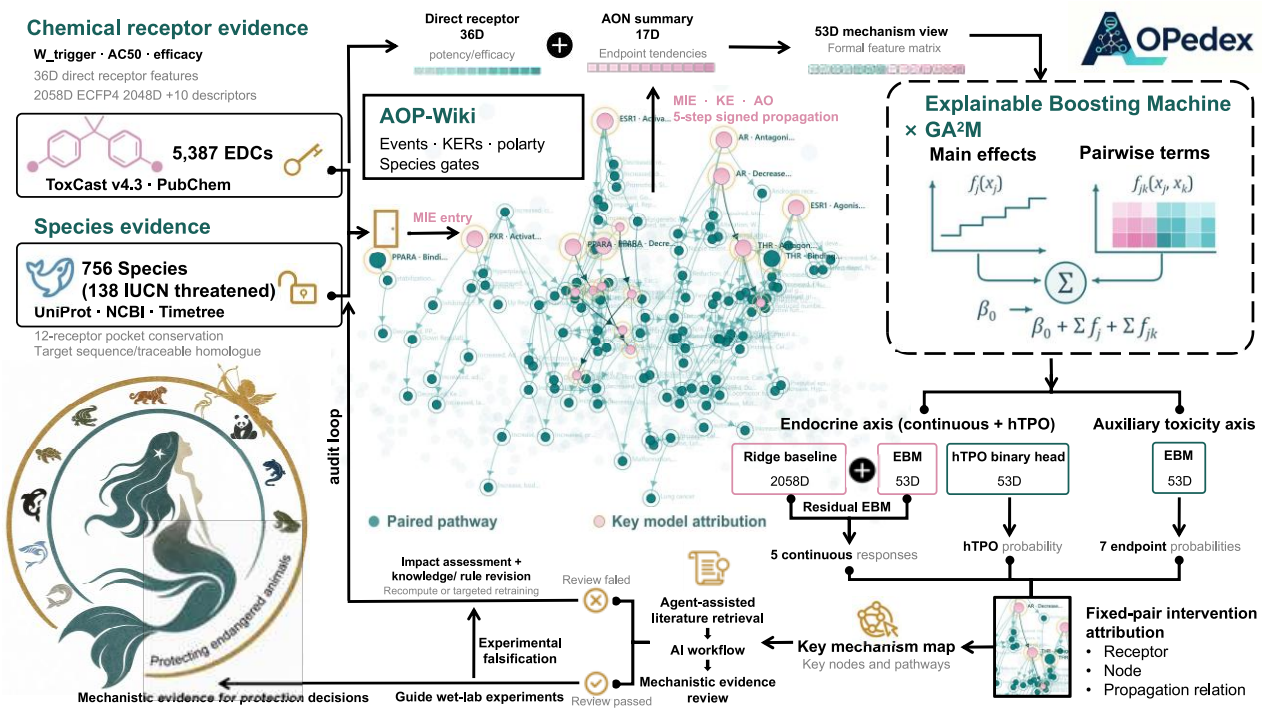


Figure 1 | Overview of the AOPedex framework. Compound receptor activity spectra (12 receptors, 5,387 active compounds and 19,557 compound–receptor pairs) are multiplied with species receptor binding-pocket conservation (756 species × 12 receptors) at the receptor molecular initiating event (MIE), propagated directionally along the 629-node AOP mechanistic network (AON) with 541 propagatable edges, and compressed into a 17-dimensional mechanistic signal. The mechanistic signal and the three receptor channels form the 53-dimensional mechanistic view, which together with the 2,058-dimensional structural core constitutes the supervised input, read out in parallel by 13 task-specific supervised heads over 8,806 modelable compounds; endocrine continuous heads use a ridge structural baseline plus EBM mechanistic residual construction. Both the construction and the propagation of the mechanistic network are deterministic computations without fitted parameters. Key model attribution is generated through receptor, node and propagation-edge interventions under fixed pairings, then projected onto the mechanistic map after stability screening; the audit loop checks predictions against evidence, with literature retrieval assisted by an LLM agent and manually reviewed. Items that fail enter knowledge or rule revision and recomputation, while validated targets point to wet-lab experiments, whose results flow back as new evidence.

Compound activity and species conservation meet at the receptor MIE, and the resulting signal propagates directionally along the AON into a 17-dimensional mechanistic signal (AON17), with the receptor activity spectrum retained alongside in three 12-dimensional channels covering direction, potency and efficacy; together these form the 53-dimensional mechanistic view. Each compound is propagated twice, under the human and zebrafish reference parameterisations in turn, and the entire computation is deterministic with no

fitted parameters (Methods). Readout is organised along six axes. The five continuous endpoints define the supervised pathway layer, with the binary hTPO head completing the thyroid readout.

A unified chemical catalogue merges two label sources under DTXSID keys, comprising 9,458 compounds. Of these, 8,806 have a complete structural core and mechanistic view, defining the training and inference scope of the 13 supervised heads; 5,387 carry 12-receptor activity records (19,557 compound–receptor pairs), defining the scope of mechanistic injection. The model covers 756 genus-level representative species; when a species lacks its own receptor sequences, estimates come from genus-level proxies or receptor-stratified k-NN close relatives, with source tags retained (Methods).

The supervised layer is organised into 13 task-specific heads; the endocrine continuous heads output potency and the rest output probabilities. High-throughput labels for general-toxicity endpoints are binary activity calls, whereas endocrine endpoints record continuous variation in potency and hormone levels; binarising both families alike would discard the quantitative information carried by concentration–response relationships. The general-toxicity and continuous endocrine heads use the AON propagation structure with separate task-specific readouts; hTPO uses a structural binary head.

General-toxicity heads and endocrine heads measure two response families on the same catalogue, the former capturing general injury and whole-organism toxicity phenotypes and the latter specific perturbations along the four EATS axes. Combinations of high and low values on the two axes define four candidate priority regions (Supplementary Table 8). This separation implements OECD GD 150 by evaluating EATS endocrine activity independently from general toxicity and non-specific stress[42].

On the classification side, the primary head is an explainable boosting machine (EBM), whose additive shape functions retain per-feature contributions and per-sample attribution[43,44], alongside logistic regression (LR) as the linear reference, XGBoost (XGB) for tree ensembles and TabPFN as the strong few-shot black box[45]. The endocrine continuous side adopts a residual structure in which a ridge regression backbone fits structure-based potency and an EBM learns the out-of-fold (OOF) residuals on the mechanistic view; three residual controls share this construction (Methods). Both sides share a unified DTXSID/scaffold outer split with fixed or default hyperparameters, so comparisons focus on the readout itself.

Predictive performance and readout quality of the dual observation axes

With the framework in place, we ask how well the two observation axes predict, supervising all 13 endpoints under one outer split (Methods). Seven classification endpoints on the general-toxicity axis output probabilities, five continuous endpoints on the endocrine axis output potency, and the binary hTPO endpoint outputs inhibition probability. The two axes answer different regulatory questions. The former supports hazard identification from general injury signals, while the latter retains potency differences or inhibition probabilities according to the original measurement type.

On the general-toxicity classification side, the 53-dimensional EBM reaches a macro ROC-AUC of 0.810, with five of the seven endpoints above 0.8; TabPFN achieves 0.812, a paired difference of −0.0016 (95% confidence interval crossing zero), so the two models discriminate almost equally. On Morphology, the endpoint with the fewest samples, the EBM overtakes TabPFN (0.698 versus 0.661, paired difference +0.037, confidence interval not crossing zero). In Figure 2a the AUC bands of the two models sit close together, indicating limited discrimination advantage; on the Morphology slice the EBM band stays above that of TabPFN in the low-to-mid false-positive range, making the reversal clear. On the thyroid binary endpoint hTPO (n = 288), the paired OOF ROC-AUC of the EBM and TabPFN is 0.741 versus 0.742 (paired difference −0.001, 95% confidence interval crossing zero).

Figure 2

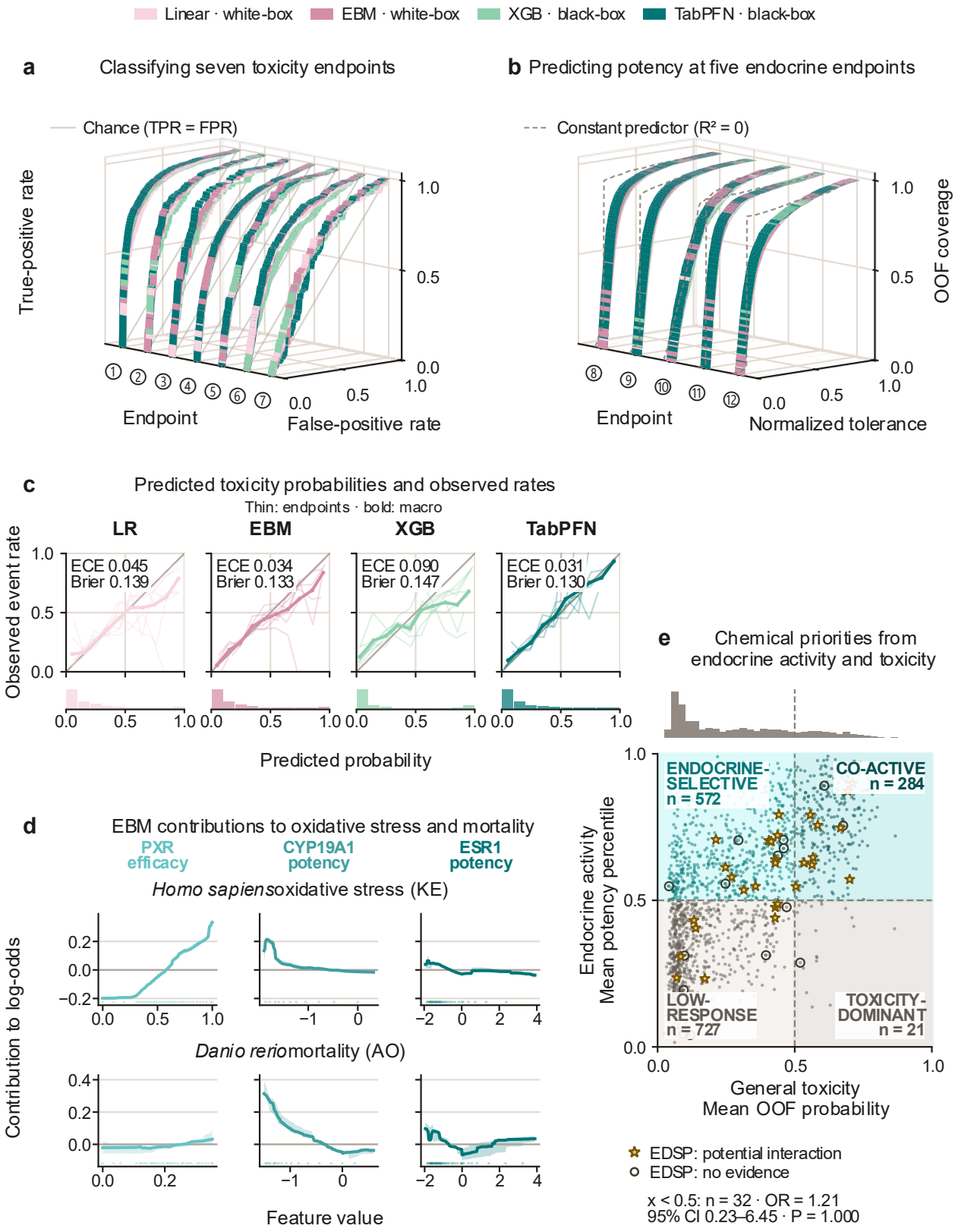


Figure 2 | Discrimination, calibration, mechanistic interpretation and compound-level convergence of the dual observation axes. a, Three-dimensional ROC bands of the four models under five-fold OOF, endpoints ordered by four-model mean AUC: ① DNA damage; ② teratogenicity score; ③ death; ④ oxidative stress; ⑤ activity; ⑥ hypoxia; ⑦ morphology. Bands of one endpoint are coplanar; the uppermost at each false-positive rate is coloured; exactly coincident bands take the EBM colour. b, REC bands of the endocrine residual models, endpoints ordered by four-model mean R^2: ⑧ ER agonism; ⑨ AR antagonism; ⑩ H295R steroidogenesis; ⑪ AR agonism; ⑫ ER antagonism.

Tolerance is normalised by ε99, the 99th percentile of pooled absolute error across the four models (0.28, 0.37, 20, 4.1 and 4.0 for ⑧–⑫); vertical axis, OOF coverage within tolerance; dashed line, constant predictor ($R^2 = 0$). Colours denote the linear model (LR in a and c, ridge regression in b), EBM, XGB and TabPFN. c, Calibration curves; thin lines, individual endpoints; thick line, macro average with equal endpoint weights; grey line, ideal calibration; predicted-probability distributions below, macro ECE and Brier annotated. d, EBM shape functions for the human KE oxidative stress and the zebrafish AO death endpoints; curves, five-fold medians; shading, interquartile range across folds; bottom rug, feature distribution; contributions are centred on the training-fold mean. e, Dual-axis OOF convergence of the 1,604 compounds; horizontal axis, endpoint mean of general-toxicity probabilities; vertical axis, mean of within-endpoint percentile ranks of endocrine potency. The 0.5 dashed lines define four candidate priority regions (counts annotated; these are priority regions, not new biological categories). EPA EDSP Tier 1 conclusions (external evidence, not involved in modelling) are overlaid: gold stars, at least one E/A/T or steroidogenesis line of evidence supporting potential interaction; grey open circles, none. For the 32 compounds with horizontal axis < 0.5 and unambiguous conclusions, the two-sided Fisher test gives a sample odds ratio of 1.22 (exact 95% CI 0.23–6.45, P = 1.000). Region denominators, aggregation and the external test are defined in Methods.

On the endocrine continuous side, the residual EBM achieves a macro R^2 of 0.2870 and a macro Spearman rank correlation of 0.4068. Relative to residual TabPFN under the same residualisation scheme, macro R^2 is lower by 0.0241 (95% CI −0.0361 to −0.0125), while macro Spearman differs by +0.0051 (95% CI −0.0013 to +0.0113). The point estimates of the ranking differences are positive for all five endpoints; endpoint-by-endpoint comparisons are shown in Figure 2b.

Output quality and mechanistic auditability are detailed in Supplementary Table 6 and Figure 2c–d. Agreement among probability, error and extrapolation confidence is assessed with ECE, Brier score, ENCE and conservativeness distance (Methods)[46], and systematic bias with the calibration slope. Without post hoc recalibration, the macro ECE10 and Brier score of the EBM are both close to those of TabPFN; the per-endpoint calibration slopes of the residual EBM range from 0.80 to 1.04, consistent with the calibration curves (Figure 2c). ENCE computed across quartiles of mechanistic signal magnitude shows that error peaks in the third quartile and falls back in the fourth for most endpoints, whereas Hypoxia error increases with signal. Figure 2d shows the EBM shape functions for the human KE oxidative stress and the zebrafish AO death endpoints; the contributions of PXR, CYP19A1 and ESR1 shift with endpoint and reference species, decomposing similar overall AUCs into distinct receptor- and pathway-level bases.

Bridging and fusion experiments test where the two axes meet (Methods). A bridging model that feeds general-toxicity probabilities into the endocrine side after the mechanistic representation failed the pre-specified global confirmation criterion, and cross-head late fusion brought no stable gains eligible for the main model. This agrees with the observation of Zhang et al.[47] that big-data-driven AOP construction tends to compress chemical-specific initiating mechanisms into generic stress responses; the oxidative stress, DNA damage and hypoxia KE endpoints on our general-toxicity axis are the representative pathways they list.

Dual-axis supervision trains and tests generic stress separately from EATS-specific perturbation, and the two signal families meet at the mechanistic representation layer.

For the 1,604 compounds under both forms of supervision, compound-level OOF aggregation forms a continuous dual-axis readout (Figure 2e; Methods). The general-toxicity coordinate is the endpoint mean of unrecalibrated EBM OOF probabilities, and the endocrine coordinate is the mean of within-endpoint percentile ranks of residual-EBM potency predictions. A pre-specified probability threshold of 0.5 on the horizontal axis and the cohort median of the supervised compounds on the vertical axis divide the plane into four candidate priority regions. The region of low general toxicity and high endocrine activity contains 572 compounds, candidates for endocrine-selective perturbation; the region of high general toxicity and high endocrine activity contains 284, where general injury and endocrine action coexist. EPA EDSP Tier 1 conclusions are overlaid as an independent external evidence layer. The external comparison includes 32 compounds and is treated as exploratory (Methods). No difference in the proportion of potential interaction evidence was detected between the high- and low-endocrine layers. The two axes therefore yield a compound-level candidate list in which endocrine activity can be examined alongside general toxicity and independent EDSP evidence.

Readout direction is further checked against a pre-specified control panel. It comprises 20 positive controls, 19 of them training-observed labels, and 10 abstaining negative controls; the abstaining controls provide the main independent negative information. Predicted directions agree for all 20 positive controls, with entry-receptor hits for 19 of 20; the sole exception is fulvestrant, whose measured THR potency exceeds that of ESR1. Of the 10 abstaining negative controls, eight are unflagged across all five axes; 2,6-di-tert-butyl-4-ethylphenol shows a weak borderline ER signal, and FD&C Yellow 6 gives a structure-driven false positive. Control-by-control results and attribution checks are listed in Supplementary Table 9. The panel therefore provides concrete checks of direction, receptor attribution and abstention, while the borderline and structure-driven signals give specific cases for mechanistic audit.

Mechanistic dissection and ablation validation

Beyond predictive performance, we ask what information the predictions rely on and whether the orientation and node semantics of the AON are preserved. A set of mechanistic coordinates independent of predictive performance answers this (Methods). For the true AON, direction concordance under the human and zebrafish

references is 0.4777 and 0.4993, and signed rank correlation is 0.5441 and 0.5334. Direction concordance counts outputs without signal as discordant; when the 59 anchor pairs are restricted to those with directed outputs, concordance rises to 0.892 and 0.875. Anchors without signal identify downstream registrations that can be examined to extend AON coverage.

Beyond the literature anchors, orientation is independently checked against regulatory evidence. The 20 pre-specified anchor pairs are drawn from regulatory evidence and primary literature, not routed through AOP-Wiki or ToxCast labels (Methods; pair-by-pair list in Supplementary Table 12). Anchor independence is limited to the evidence source; compound identities still overlap the training set. Directions agree for 12 of 13 positive anchors and for 0 of 7 negative anchors. Disagreements are decomposed into three categories — differences in measurement definition, threshold-boundary cases and genuine extrapolation conflicts — and a disagreement may fall into more than one. These anchors provide pair-level evidence for checking mechanism direction and for classifying the sources of disagreement.

The mechanistic view is generated with task-specific masking, blocking the leakage of label-provenance information into the feature side (Supplementary Note 1). The direction lexicon, topology and node semantics all come from manually curated priors, and whether they genuinely shape the predictions is tested by controlled perturbations. Six controlled perturbations are constructed alongside the true signed AON (A0) (Methods), and evaluation follows the 59 literature orientation anchors. Direction concordance and signed rank correlation measure the preservation of overall orientation proportion and receptor-level response ordering, respectively.

The mechanistic properties of each perturbation are characterised by the two metrics (Figure 3a). Polarity randomisation (A3) is the most destructive, dropping the two metrics to 0.2195/0.2293 and 0.0270/0.0224, and ten seeds reproduce collapses of similar magnitude under both references. Matched rewiring (A2) and node-semantics permutation (A4) maintain direction concordance close to the true AON, while signed rank correlation falls to 0.38–0.46. Overall orientation proportion is easy to preserve, but receptor-level fine ordering depends on the true topology and node semantics; the advantage of the true AON requires both metrics to hold.

Figure 3

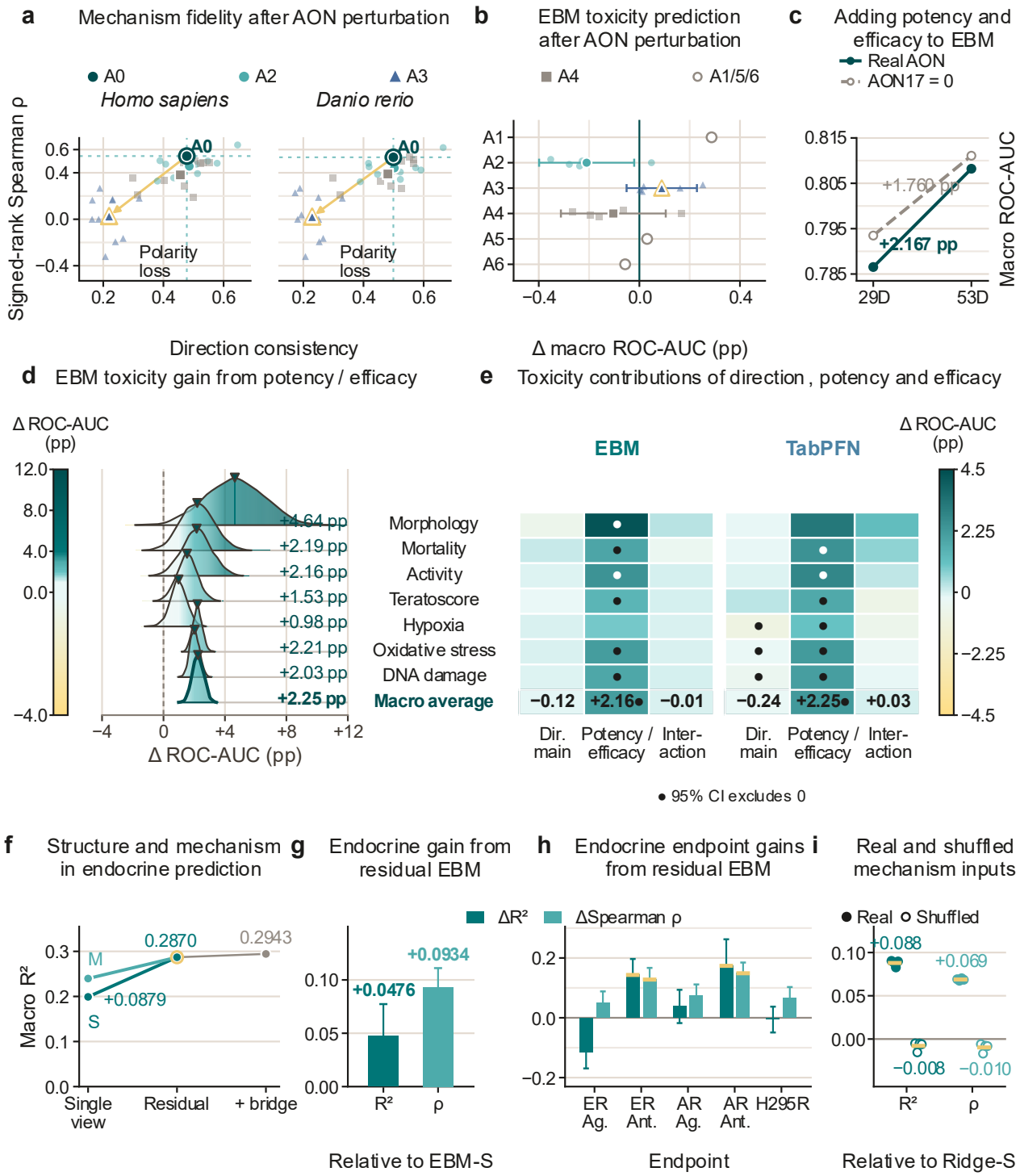


Figure 3 | AON mechanistic fidelity, pharmacological-information increments and structure–mechanism complementarity. a, Direction concordance and signed rank correlation under the human (Homo sapiens) and zebrafish (Danio rerio) references; small points, mechanistic seeds; large points, means; arrows connect the A0 and A3 means. A0, true AON; A1, no propagation; A2, matched rewiring; A3, polarity randomisation; A4, node-semantics permutation; A5, species gating removed; A6, six-step propagation; A2–A4 each have 10 mechanistic seeds. b, Macro ROC-AUC differences of A1–A6 relative to A0; A2–A4, means over the five seeds with seed scatter and 95% t intervals; A1, A5 and A6 are deterministic controls; pp, percentage points. c, The 29D→53D representation bridge under the true AON (solid) and zeroed AON17 (dashed), first seed of one outer split. d, Per-endpoint paired ROC-AUC differences (53D − 29D) for the EBM; inverted triangles, point estimates; five-seed mean predictions resampled by paired bootstrap (B = 2,000). e, Direction-coding main effects, potency/efficacy main effects and interaction terms of the white-box EBM and

black-box TabPFN, under one outer split and the same seed; ●, 95% interval not crossing zero; values annotated only in the macro-average row. d and e share endpoint names in the middle; colour scales span −4 to +12 pp (ridgelines, left) and −4.5 to +4.5 pp (heatmaps, right); macro averages weight the seven endpoints equally. f, Macro R^2 of the structural ridge (S, 2,058D structure), the direct mechanistic EBM (M, task-masked 53D mechanistic features), the residual EBM and the probability-bridging arm; connected differences use metric means over four confirmation seeds. g, h, ΔR^2 and ΔSpearman ρ of the residual EBM against the pure structural EBM (EBM-S), macro and per endpoint; bars, paired scaffold-cluster bootstrap with 95% intervals (B = 5,000); Ag., agonist; Ant., antagonist; gold bar ends, the dual-metric gains of the ER/AR antagonist endpoints. i, Macro increment of true versus jointly permuted mechanistic inputs over the ridge backbone; filled points, true inputs; open points, permuted inputs (four confirmation seeds each); gold ticks, means. Macro results in f–i weight the five continuous endpoints equally.

The same perturbations induce only narrow changes on the prediction side (Figure 3b). Macro ROC-AUC changes for the six controls relative to A0 all lie within ±0.3 percentage points. The no-propagation arm A1 differs by +0.19 percentage points in the macro (95% CI −0.98 to +0.57), with per-endpoint differences between −0.71 and +0.55 percentage points and intervals all crossing zero (Supplementary Information). A3, the largest mechanistic displacement, barely moves on the performance axis. ROC-AUC and mechanistic fidelity measure different properties of a model. A classifier can maintain discrimination from multiple correlated signals, whereas the AON audit tests whether those signals still correspond to the orientation, receptor ordering and node semantics of the established mechanistic reference.

The complementarity of mechanistic context and pharmacological information is compared in a representation bridge, expanding the features from 29 to 53 dimensions (Figure 3c). Adding potency and efficacy raises macro ROC-AUC by 2.167 percentage points under the true AON and by 1.760 percentage points under zeroed AON17, with a difference interaction of +0.406 percentage points. Pharmacological strength gives a stable increment under both conditions, and the true mechanistic context further amplifies it.

Under the unified scaffold outer split, the macro ROC-AUC increment from 29 to 53 dimensions is +0.02249 for the EBM (95% CI 0.01553–0.02955) and +0.01895 for TabPFN (95% CI 0.01248–0.02532); per-endpoint paired differences are shown in Figure 3d. All seven point estimates are positive for the EBM, six with intervals not crossing zero. Morphology has the widest distribution and the highest point estimate (+4.64 percentage points); this composite phenotype, which summarises 17 morphological components, is the most sensitive to pharmacological-strength information. Hypoxia is the only endpoint whose interval crosses zero. Its mechanistic coverage has a gap, as the current AOP-Wiki subnetwork lacks a qualified directed path from the receptor MIE to HIF, and AON17 is identically zero for this endpoint. For TabPFN, all endpoints other than Morphology have intervals not crossing zero. The magnitude of the increment thus reflects both representational capacity and knowledge coverage.

Attribution over the four representation cells locates the increment in the potency/efficacy information block (Figure 3e). The potency/efficacy macro main effects are +2.16 and +2.25 percentage points for the EBM and TabPFN, respectively. With both direction-coding intervals crossing zero, the matching EBM and TabPFN effects locate the representation gain in pharmacological strength.

The ablation logic extends to the endocrine continuous side. The residual EBM merges a structural arm and a mechanistic arm (Methods). The structural ridge and the direct mechanistic EBM achieve macro R^2 of 0.1990 and 0.2397, and the residual EBM raises this to 0.2870; relative to the direct mechanistic readout, the residual model gains +0.0484 in macro ΔR^2 (95% CI 0.0171–0.0724) and +0.0271 in macro ΔSpearman (95% CI 0.0073–0.0462) (Figure 3f). The supervised gain concentrates in the combination of structural and task-specific mechanistic features.

The second reference is a purely structural EBM, which tests whether the mechanistic residual exceeds the nonlinear capacity of the structural white box itself (Figure 3g). Relative to the pure structural EBM, the residual EBM gains +0.0476 in macro ΔR^2 (95% CI 0.0178–0.0773) and +0.0934 in macro ΔSpearman (95% CI 0.0752–0.1111), with all four confirmation seeds positive in direction. The nonlinearity of the pure structural model cannot account for the final gain, and the supplementary information in the combined receptor-pharmacology and AON view stands out most in the ranking metrics.

The division of labour varies by endpoint (Figure 3h). The ΔSpearman 95% CIs for all five endpoints lie above zero; within-endpoint potency ordering improves for the ER, AR and steroidogenesis response families, and the macro increment is jointly supported by multiple endpoints. The two antagonist endpoints show the most complete gains, where task-specific mechanistic information supplements structural prediction most strongly. The ER agonism endpoint divides the labour differently, with the pure structural EBM carrying higher R^2 ($\Delta = -0.1152$) and the residual EBM higher Spearman ($\Delta = +0.0508$); strong structure–activity relationships carry the amplitude fitting, and the mechanistic residual improves potency ordering. The shared gains of AR agonism and H295R likewise concentrate in ranking metrics.

The information source of the increment is examined with permutation controls (Figure 3i). True mechanistic inputs give macro ΔR^2 and ΔSpearman of +0.0879 and +0.0688 over the ridge structural backbone; after shuffling the compound-to-feature correspondence while keeping the number of mechanistic columns unchanged, the two increments fall back to −0.0079 and −0.0097, and the two groups separate cleanly on

either side of zero. These controls show that the supervised gain depends on the measured compound-to-mechanism correspondence.

Nested Ridge comparisons isolated the contribution of receptor measurements across five endocrine tasks. Adding receptor inputs raised task-macro R^2 by 0.0946 (paired scaffold-bootstrap 95% interval 0.0676–0.1192) and macro Spearman ρ by 0.0739 (0.0570–0.0915). Both metrics improved for every task, and macro gains were positive at every split seed (Supplementary Table 7). Receptor measurements thus supplied consistent endocrine information beyond chemical structure.

The division of labour on the ER agonism endpoint, with structure dominating amplitude and mechanism dominating ranking, suggests that structural contributions depend on molecular context. We compare the purely additive E0 model with E8, which allows up to eight second-order interactions. E8 brings a ΔR^2 of +0.1878 (95% CI 0.124–0.268), the interaction family reproduces across folds and maps back to bits of the extended-connectivity fingerprint of diameter 4 (ECFP4), and its stability and error stratification are checked in Supplementary Note 3. Native terms are then mapped to real atomic environments and read at three levels. Main-effect functions give the response contribution of each encoded bit state, interaction functions give how that contribution moves with the state of a background descriptor, and the additive form decomposes every prediction term by term into concrete structural factors (Figure 4).

Figure 4

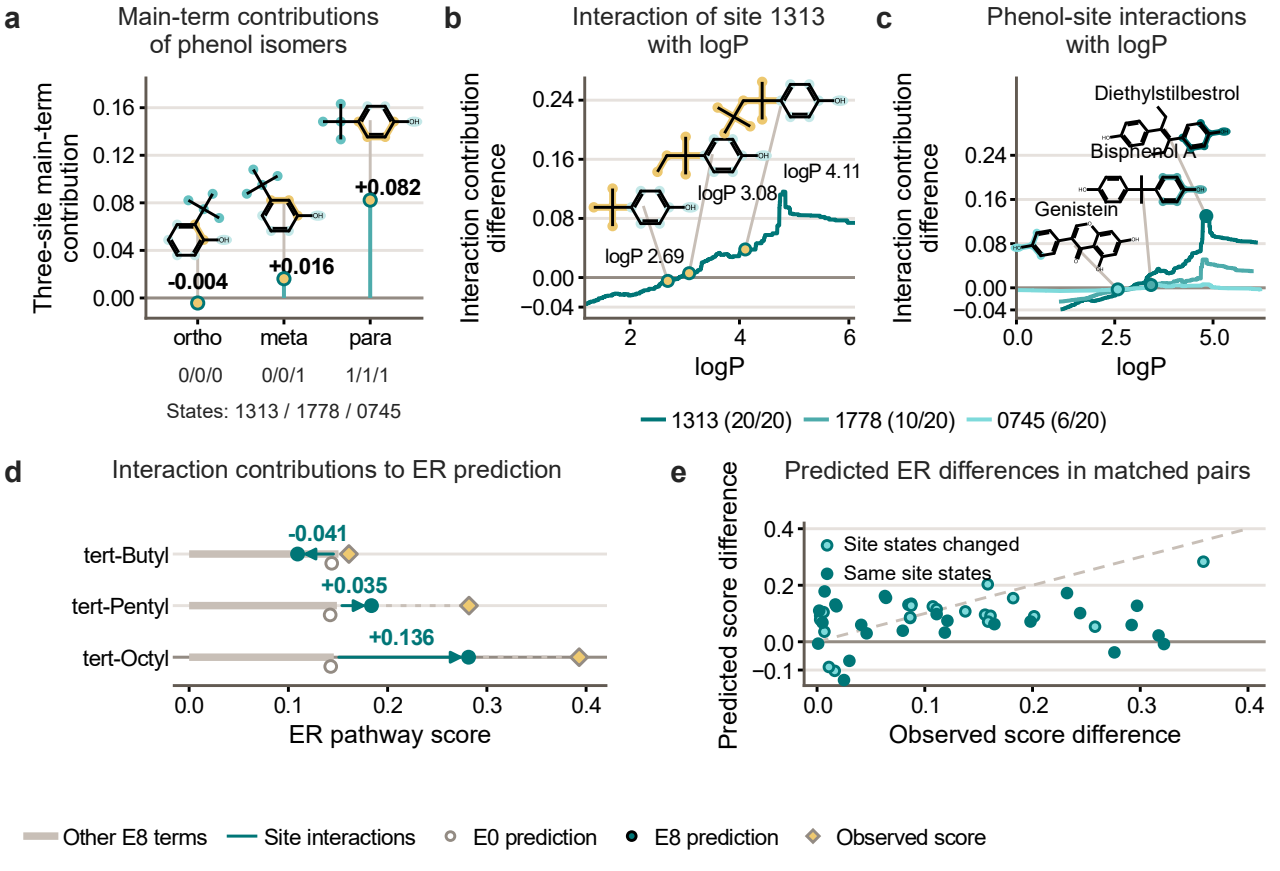


Figure 4 | Local discrimination and conditional interactions in the structural white box of ER agonists. a, Ortho-, meta- and para-tert-butylphenol and the encodings of bits 1313, 1778 and 0745; the vertical axis is the median across 20 confirmation models of the sum of the three main-effect terms; the three isomers have identical molecular weight and logP. b, Difference of logP interaction slices with bit 1313 present versus absent; curves are pointwise medians over 20 models. c, Interaction-slice differences of the three target bits with logP, averaged over all confirmation models; unselected interactions are set to zero, and legend parentheses give selection counts; highlighted atoms come from the actually decoded bit environments. Curves in b and c are restricted to the central 95% of logP among bit-positive compounds. d, Out-of-fold contribution decomposition for para-alkylphenols; grey lines are the sum of the remaining E8 terms, teal arrows are the target-bit interaction corrections, filled teal points, open grey points and yellow diamonds are E8 predictions, E0 predictions and experimental scores, and values are out-of-fold means over four confirmation seeds. e, Experimental and predicted differences for 48 pairs of real structural counterparts, oriented as the higher experimental score minus the lower; dark teal indicates unchanged target bits, light teal indicates at least one changed bit, and the dashed diagonal is exact amplitude agreement. Contribution quantities in a–d are in units of the ER pathway score.

Positional discrimination comes first. Ortho-, meta- and para-tert-butylphenol share molecular weight and logP and differ only in the encoding at bits 1313, 1778 and 0745 (000, 001 and 111), yet the summed main-effect contribution distinguishes their substitution positions (Figure 4a). The same bit is background-dependent, its logP interaction-slice difference rising from −0.0045 in the tert-butyl instance to +0.0380 in the tert-octyl instance and turning the same local environment from slightly negative to positive as the alkyl group grows (Figure 4b). The three bit–logP interactions are selected by 20, 10 and 6 of the confirmation models, with bit 1313 the most stable interpretable entry (Figure 4c).

These context-dependent contributions are also evident in out-of-fold predictions (Figure 4d). The remaining E8 terms and the corrections of the three target-bit interactions add exactly to the output; across the tert-butyl, tert-amyl and tert-octyl series the corrections move with the molecular background while E0 predictions cluster at 0.142–0.144. Using a shared-compound design, we tested 48 pairs of real structural counterparts (Figure 4e); the reading agrees with experiment in direction for 41 pairs, and the 25 pairs that keep all three target bits fixed raise direction agreement from 48% under E0 to 80% under E8. Across these structural counterparts, the interaction terms improve directional agreement with experiment while preserving an exact decomposition of each prediction.

Stability, species resolution and extrapolation error in cross-species transfer

The OECD Users' Handbook requires every assessment to state its taxonomic applicability domain[48], and manual judgement cannot cover thousands of species. Haigis et al.[34] note that the cross-species applicability of downstream KEs and AOs in an AOP is usually weaker than that of the MIE. We therefore measure input estimation error and changes in the mechanistic discriminative readout under cross-species transfer separately.

Using the same set of measured zebrafish labels, we compare fish-source and human-source parameterisations of the mechanistic signal. Across the four zebrafish endpoints, both parameterisations reach a macro ROC-AUC of 0.7298, and the parameterisation difference Δ (fish − human) is −0.000069 (−0.000488 pooled over five seeds). At the per-compound level, after registered flagship compounds switch from the human to the fish reference, the median prediction shift is 0.94% of the label standard deviation (p90 4.8%), none of the 1,625 compound × endpoint cells flips direction, and the pooled Spearman correlation of compound rankings is 0.9987 (Methods). The shift is carried by the direct receptor injection; with propagation switched off the displacement shrinks, and on the H295R endpoint the full-propagation displacement is about 1.9 times that of the no-propagation control. Of all 9,458 compounds in the catalogue, 49.5% have non-zero receptor injection, and the direction-flip rate in this subset is 4.6%, concentrated in near-zero-score compounds. Reference-species switching therefore preserves compound rankings and discriminative readout across the measured zebrafish endpoints.

The mammal-to-fish gap shows the opposite pattern at the sequence level and at the mechanistic level (Figure 5a, b). From within-class holdout to cross-class extrapolation, whole-sequence conservation error rises from

0.037 to 0.282 (7.67-fold) and pocket error from 0.055 to 0.187 (3.43-fold); the pocket reduces the cross-class penalty by 0.113 (95% CI 0.025–0.201), and mechanistic discrimination is barely affected.

Figure 5

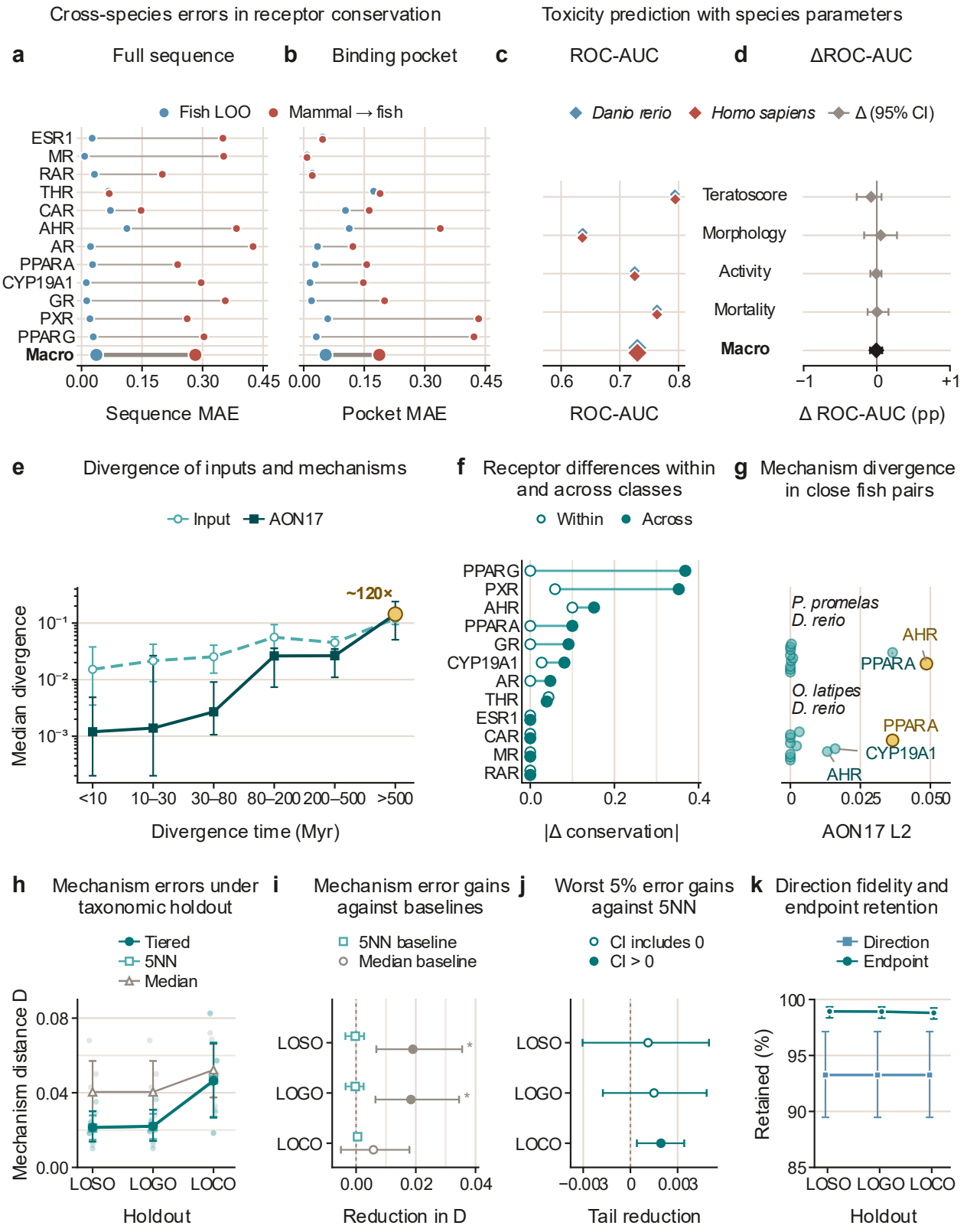


Figure 5 | Input cost, mechanistic resolution and holdout fidelity of cross-taxon extrapolation. a, b, Conservation MAE of whole sequence and binding pocket, for within-class leave-one-out in fish (Fish LOO) and mammal-to-fish extrapolation;

grey lines connect the same receptor; large points, receptor-equal-weight macro means; a and b share receptor names on the left. c, Five-fold ROC-AUC of fish-source and human-source parameterisations on the four zebrafish endpoints; large diamonds, macro means. d, Fish-source minus human-source ROC-AUC, paired bootstrap 95% CI (B = 10,000); pp, percentage points; c and d share endpoint names in the middle. e, Input differences (Input) and AON17 mechanistic distance for species pairs in six phylogenetic-distance bins: 16 pairs per bin, 24 receptor probes per pair. Points, medians; intervals, interquartile ranges; vertical axis, logarithmic; zero lower quartiles shown at the display floor 2×10^{-4}; gold points, farthest bin; ~120-fold, farthest-to-nearest bin ratio. f, Effective conservation difference of each receptor within class (open) and across classes (solid). g, Median AON17 L2 difference of receptor-dominant compound subsets: fathead minnow (Pimephales promelas) and medaka (Oryzias latipes) each against zebrafish (Danio rerio); gold points, each pair's maximum difference; leader lines mark receptors. Annotated values: fathead minnow 0.0486 (AHR, n = 163), 0.0365 (PPARA, n = 37); medaka 0.0365 (PPARA, n = 37), 0.0160 (CYP19A1, n = 110), 0.0132 (AHR, n = 163). h, Class-equal-weight mechanistic distance D under species- (LOSO), genus- (LOGO) and class-level (LOCO) holdout: Tiered, receptor-stratified rule; 5NN, uniform 5-nearest-neighbour; Median, global median. Light points, per-class estimates; connecting lines, macro estimates; error bars, 95% CI. i, Reduction in D of the stratified rule relative to both baselines (positive, improvement); solid points and *, two-sided paired bootstrap $P < 0.05$; open points below the threshold. j, As i but for the worst 5% of compounds per species, relative to 5NN; open symbols, 95% CI includes zero; solid symbols, CI entirely above zero. k, Direction agreement and dominant-endpoint retention. h–k, stratified bootstrap (B = 2,000) resampled by class → genus → species → compound; error bars, 95% CI.

Mechanistic divergence amplifies stepwise with phylogenetic distance (Figure 5e). AON17 differences between species pairs in six phylogenetic-distance bins, probed across 24 receptor probes, stay near 10^{-3} in median for the three nearest bins, with jumps at 80 Myr and 500 Myr; the nearest bin (0.00120) and the farthest bin (0.14393) differ by about 120-fold, and Spearman's ρ between distance and difference is 0.702. Conservation-input differences move in the same direction but more gently (the farthest bin exceeds the nearest by only 7.9-fold), so the amplification lies mainly in the propagation structure rather than in the inputs.

The amplification is uneven across receptors (Figure 5f). ESR1, MR, RAR and CAR show the smallest within- versus cross-class differences, forming a conserved base shared across species; PPARG (0.368) and PXR (0.353) show the largest. THR is the only receptor whose cross-class difference does not exceed its within-class difference (0.039 versus 0.043). The divergence of the two close-relative pairs is contributed by different receptor ports and varies with the species pair (Figure 5g). The proportion of sequences drawn from the species' own genome correlates significantly with neither phylogenetic distance ($\rho = -0.134$) nor mechanistic difference ($\rho = 0.043$); with VTG gating switched off, the distance–mechanism correlation persists ($\rho = 0.68$), and the separation of the farthest bin is carried jointly by continuous conservation differences and gating.

Holdout protocols give the quantitative loss. LOSO, LOGO and LOCO are executed over 904 species, 639 genera and five classes, sharing 220 compounds (Methods). Under class-stratified rules the class-equal-weight mechanistic distances are 0.02142, 0.02205 and 0.04643, and LOCO is 2.17 times LOSO ($p = 0.001$). For average error, the stratified rule shows no significant difference from the uniform 5-NN and is significantly better than the global median under LOSO and LOGO ($p = 0.001$), while under LOCO it ties (CI crossing

zero). The severe-error tail differs. Under LOCO, the mechanistic distance of the worst 5% of compounds per species drops by 0.00194 (95% CI 0.00041–0.00342, p = 0.011), whereas the corresponding intervals for LOSO and LOGO cross zero. Direction agreement remains about 0.93 and dominant-endpoint retention about 0.99 across the holdouts (Figure 5k), showing that the stratified rule preserves the mechanistic readout as quantitative error increases.

The error–distance relationship of each receptor determines the extrapolation route. Strong-signal receptors amplify 5.6–13.5-fold from close relatives to cross-class and rise monotonically with distance; mid-signal receptors amplify 2.0–7.1-fold; weak-signal receptors saturate within class and no longer amplify, consistent with the conclusion of Tong et al.[49] that confidence declines with extrapolation distance. We stratify receptors by Mantel r, extrapolating those with $r \geq 0.4$ by cross-genus k-NN and falling back to the global median for those with $r < 0.15$; the stratification statistics follow Blomberg et al.[50], and the basis for using a generic model for weak signals is given by Shang et al.[51]. 92.8% of species have at least five close-relative references; replacing a target species' conservation with that of close relatives changes the mechanistic signal by 0.005, versus 0.053 with distant relatives, an 11.4-fold difference (Wilcoxon $p = 2.96 \times 10^{-18}$, discriminative AUC 0.911), stable to compound and threshold choices.

The choice of fish reference species does not change the conclusions. With medaka (Oryzias latipes) or fathead minnow (Pimephales promelas) replacing the zebrafish reference, predictions for all 9,458 compounds on the five endocrine endpoints correlate with the zebrafish reference at Spearman 0.92–0.99, with a zero median prediction difference and direction-flip rates no higher than 5% (Supplementary Table 13). Candidate-sequence replacement in receptor pockets does not change state assignments either. After 397 cells are recomputed, none of the three states of the 756 species changes (Supplementary Table 14).

Published fish aromatase measurements[93] provided an external test of pocket-guided transfer. We withheld each species together with all its responses before fitting. In the 12 fishes with same-species sequences, CYP19 pocket conservation reduced species-macro GMFE from 2.801 to 2.072 across 21 IC50 measurements. Non-pocket identity gave 3.494 (Supplementary Table 16). These independent measurements support CYP19 pocket conservation as a sequence-level basis for aromatase extrapolation.

ECOTOX records test the close-relative strategy with holdout by species. The loss in predicting reproductive outcomes from close-relative conservation is 0.011, versus 0.019 from distant relatives, with close relatives

512 better for 70.2% of species; this test is a stricter examination of out-of-distribution species extrapolation than
513 compound holdout[52]. On the species-holdout discrimination tasks, species-specific mechanistic features
514 exceed the 12-dimensional receptor-activity baseline on endocrine-specific labels (endocrine enzyme activity
515 0.709 versus 0.703; sex hormones 0.650 versus 0.642). Full per-label comparisons and ICE-style interspecies
516 regressions are listed in Supplementary Table 17. Together, the holdouts show that close-relative conservation
517 improves reproductive-outcome prediction, while species-specific mechanistic features add discrimination on
518 endocrine-specific labels.

519 **Applicability-domain coverage across the vertebrate phylogeny and assessability of threatened**
520 **species**

521 We now ask where across the vertebrate phylogeny these transfer rules make mechanistic assessment possible.
522 Whether mechanistic knowledge transfers across species rests on the conservation of the mechanisms
523 themselves. The hypothalamic–pituitary–gonadal axis is highly conserved between fish and mammals, and so
524 is the hypothalamic–pituitary–thyroid axis between amphibians and mammals[10], while ligand-binding
525 domains of nuclear receptors are protected by purifying selection across vertebrates[53]; these conserved
526 structures are the basis for cross-species extrapolation. OECD ecotoxicology test guidelines and the US EPA
527 Endocrine Disruptor Screening Program (EDSP) likewise focus on these taxa and their EATS pathways.
528 Invertebrate endocrine regulation is taxon-specific. Insects regulate development with ecdysteroids and
529 juvenile hormone, mollusc and crustacean mechanisms are still being worked out[54], and evidence for
530 oestrogen-like effects remains insufficient[55]. The applicability domain of the model is jointly defined by core-
531 receptor confidence and endpoint applicability (Methods).

532 Building on these conserved structures, we compute the three prediction states for 756 species × 12 receptors,
533 and 209 species come out high-confidence, 505 predictable and 42 unpredictable. High-confidence or
534 predictable states cover more than 90% of species in each of the five largest sampled vertebrate classes,
535 including reptiles and amphibians (Figure 6c). Insufficient information is retained as a model output, and
536 sequence-coverage bias is presented alongside. Sequence evidence varies across classes, with some species
537 relying on close-relative proxies or default values. Cell-by-cell provenance and input-equivalence classes for
538 the 9,072 species × receptor cells are registered in Supplementary Note 5; 68.7% of cells rest on direct
539 alignment of the species' own or reference sequences, and species-specific information cells total 77.7%.

Default-filled cells account for 40.6% in cartilaginous fish, and no jawless fish reaches high confidence (0/5). These patterns could guide the selection of additional representative model species for regulatory endocrine testing, particularly in cartilaginous and jawless fishes. Building receptor-sequence and endocrine-response reference data for these lineages would support the extension of standardised assays across a wider range of vertebrates.

Figure 6

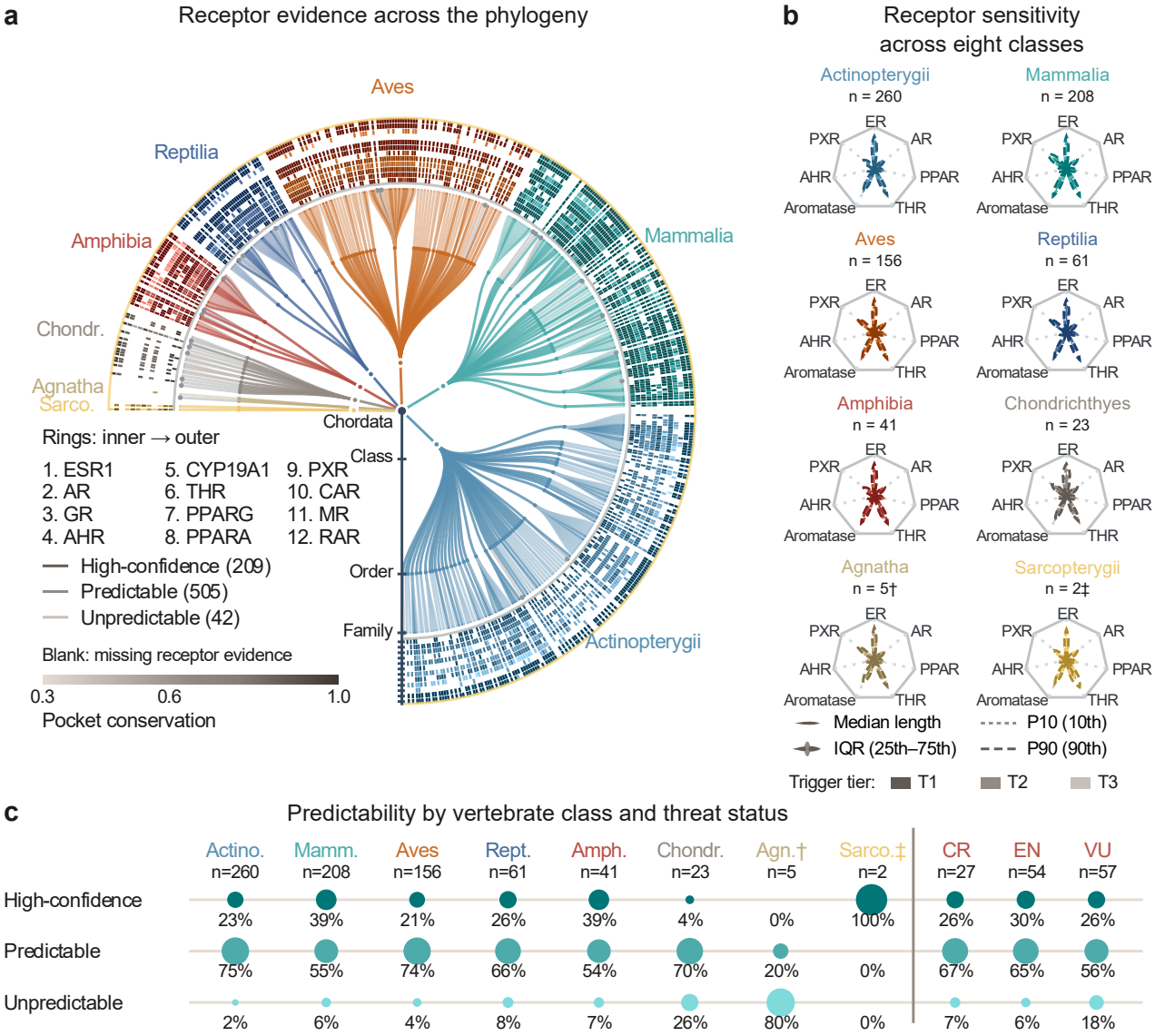


Figure 6 | Receptor-evidence coverage and predictability across the vertebrate phylogeny. a, Phylogenetic tree of the 756 species with a ring heat map of pocket conservation across the 12 receptors; colour denotes class, shading denotes pocket conservation, and blanks denote insufficient receptor evidence. Branches and species are marked as high-confidence, predictable or unpredictable. b, Petal plots of model receptor-response indices under seven fixed compound–receptor probes for the eight vertebrate classes. Each index is the absolute fixed-probe injection value multiplied by the species' registered receptor-conservation estimate; petal length denotes the class median, width the IQR, and dashed outlines the tenth and ninetieth percentiles. † and ‡ mark small-sample classes. c, Composition of the three predictability states within each class and within the IUCN CR, EN and VU threatened subsets; bubble area and labels both give within-column proportions. Class panel, n = 756; threatened subsets, n = 138.

Of the 756-species panel, 138 are IUCN threatened species[56], and 123 of them fall into high-confidence or predictable states; threat level and predictability are read out together, with mechanistic assessability shown alongside evidence reliability. Reptiles and terrestrial-phase amphibians are poorly represented in standardised wildlife testing and assessment frameworks[12], and bio-read-across evidence is likewise scarce[10]. In these taxa the model relies on MIE anchoring, the aromatase mechanistic chain and literature anchors. Pollutant exposure can alter aromatase expression in turtle embryos[57], and Cyp19a1 is an essential gene for ovary determination in reptiles[58]. An exploratory comparison uses six reptile reference species, with intervals in the Supplementary Information. Five of the six reptile reference species thus support the CYP19A1-dominant direction recovered by the model, connecting the sequence-based assessment with a conserved reproductive mechanism.

Of the 756 species, 714 reach high-confidence or predictable states, 94.4% (Figure 6a, c). Ray-finned fishes have the deepest and fullest receptor rings, jawless fishes are almost entirely blank, and unpredictable species often show local gaps in their receptor rings. The ER probe produces the largest median receptor-response index in all eight classes (0.79–0.99), with the Aromatase and THR probes next (Figure 6b). Under the fixed PXR probe, the median index is 0.671 in birds and 0.629 in mammals, compared with 0.335 in ray-finned fishes; the AHR probe also gives its lowest median index in ray-finned fishes (0.197). Among these fixed probes, PXR shows the largest between-class spread in the model index.

The three predictability states are further broken down by class and by threatened subsets, with rows as states, columns as the eight classes and three threatened subsets, and bubble area and labels both giving the within-column proportion (Figure 6c). Actinopterygians fall 75% into the predictable band; mammals and amphibians have the largest high-confidence shares (39% each); 26% of cartilaginous fish are unpredictable and 80% of jawless fish. The threatened subsets on the right are read the same way. The proportions of CR and EN species in high-confidence or predictable states are 93% and 94%, respectively, comparable to the full panel's 94.4%. VU species also retain broad coverage, with 18% classified as unpredictable. These results show that broad model applicability can be retained for threatened species even when species-specific toxicity records are absent.

Stage-by-stage white-box audit and iterative revision of the 3D mechanistic map

The same predictions also power an interactive 3D mechanistic map covering 756 species and 5,387 compounds, supporting mechanistic retrieval for 4,072,572 species–compound pairings (about 4.07 million) and linking about 17.81 million model predictions with 102.3 million attribution records. The AOPedex online database (https://aopedex.com) presents predictions, endocrine-disruption pathways and key model attribution for compound–species pairings together. The homepage shows the data scale and research background (Figure 7a, b); the interactive map contains 629 event nodes, classified as MIEs and KEs/AOs with arrows denoting relation direction, and can be filtered by receptor channel or official AOP (Figure 7c). The query page provides searchable, copyable and downloadable species and compound catalogues, supporting retrieval by scientific name, compound name or CAS number (Figure 7d). A joint query for giant panda × BPA shows pairing information, mechanistic entries and event cards on the left, and highlights the endocrine-disruption pathway and key model attribution members on the right; mechanism cards can be exported to Excel or JSON for mechanistic verification and validation-experiment design (Figure 7e). Every prediction and attribution member shown in the interface unfolds along the same white-box computation chain.

Figure 7

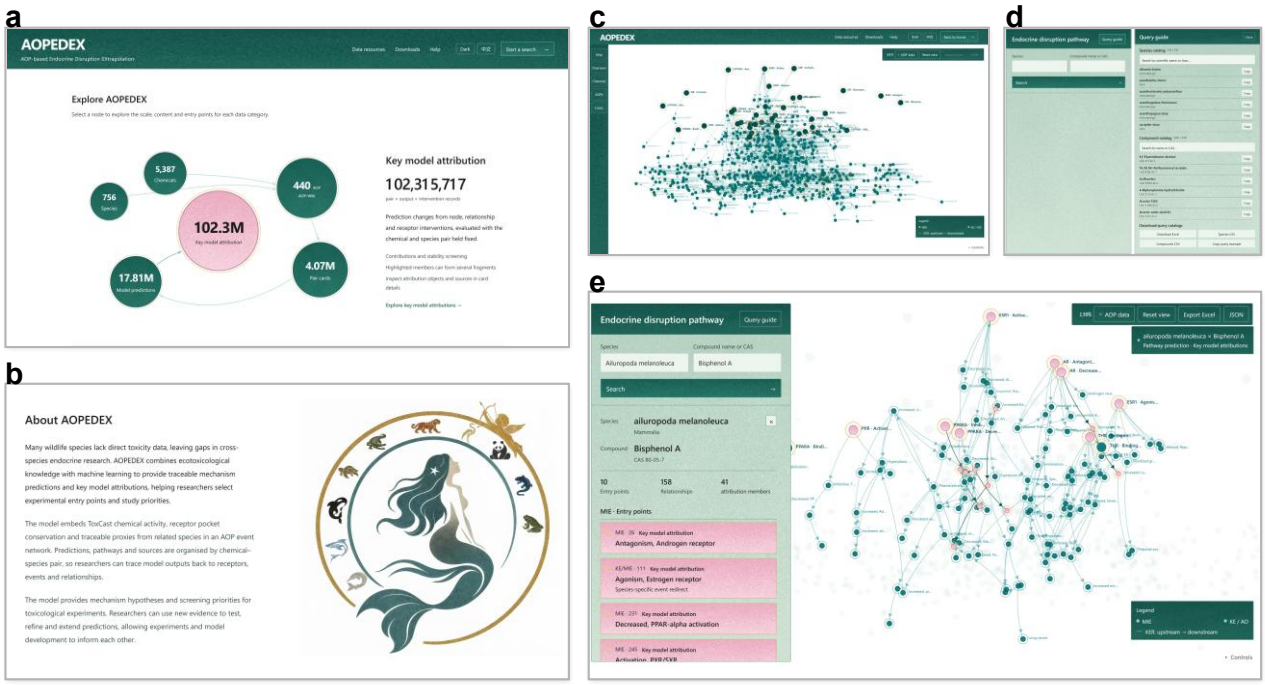


Figure 7 | Online interface of the AOPedex database and pairwise retrieval. a, Homepage data overview showing the scale of species, compounds, associated AOPs, pairing mechanism cards, model predictions and key model attribution. b, Database introduction and concept diagram, illustrating mechanistic organisation, prediction and evidence tracking under compound–species pairings. c, Interactive AOP master map showing all 629 event nodes and their relations; nodes are classified as molecular initiating events (MIEs) and key events/adverse outcomes (KEs/AOs), and arrows denote relation

direction. d, Query guide providing species and compound catalogues with name search, copying and list download. e, Joint retrieval example for giant panda (Ailuropoda melanoleuca) and bisphenol A (BPA). The left shows pairing information, entry events and mechanism cards; the right shows the endocrine-disruption pathway, with light-pink filling and gold markers highlighting key model attribution members. Key model attribution is based on prediction differences produced by node, relation or receptor interventions under fixed pairings, after stability screening; highlighted members can be distributed across multiple pathway segments.

The white box expands the prediction process into seven audit stages, each open to checking and intervention (audit loop at the bottom of Figure 1), covering compound input, receptor conservation, event entry, pathway propagation, supervised readout, intervention attribution and no-output reasons. Receptor inputs retain assay and sequence provenance, entries retain event identity and mode of action, and pathway relations retain upstream and downstream events, tissue context and stopping basis. The ridge structural base score and the EBM mechanistic residual can be rebuilt from the fitted parameters, and EBM main effects and interaction terms can be read directly, so anomalies can be traced upward to the knowledge registry and downward to specific computation terms. Key model attribution is produced by in-model intervention (Methods). Pathway, domain, contribution and five-seed stability screens then project qualified members onto the mechanistic map, with all other outputs retained in the audit record.

The white-box architecture connects feature contributions to mechanistic semantic groups, internal nodes, propagation paths and knowledge sources (Supplementary Table 23). At the feature layer, the classification-head EBM has a GA²M structure with 53 main effects and 159 second-order interaction terms. Per-feature contributions reconstruct the logit-space prediction, with a maximum identity residual of 9.6×10^{-14} across all out-of-fold predictions. After retraining from scratch with changed seeds and outer splits, pairwise Spearman correlations of global attribution vectors across the seven classification endpoints have a macro mean of 0.987 (0.973–0.994 per endpoint), with a top-10 feature Jaccard of 0.932.

The SHAP comparisons examine background-set sensitivity and agreement across models and methods. For a fixed TabPFN model, changing the Permutation SHAP background set gives a cross-background Spearman macro mean of 0.906 (0.831–0.968). On the same task and feature set, Permutation SHAP for TabPFN and TreeSHAP for XGBoost reach a global rank correlation of 0.605; this comparison changes both the fitted model and the attribution method (Supplementary Table 21). These analyses assign contributions to input features. The white-box model reads feature contributions directly from its fitted terms, while the Permutation SHAP comparison uses per-sample perturbation sampling.

Below the feature layer lies the mechanistic layer. Every dimension of the 53-dimensional mechanistic view has registered semantics — receptor direction, potency, efficacy or an AON17 event group — and the additive decomposition lets contributions be merged into mechanistic semantic groups, with interaction-term contributions split equally among member features. The four groups of AON17 state, receptor injection signal, receptor potency and receptor efficacy completely cover the prediction output for each endpoint, and the receptor-efficacy channel is the largest contribution group in all seven endpoints (31%–37%, Supplementary Table 22). Further down, mechanistic nodes themselves become intervention objects. With pairings fixed, AON nodes are clamped group by group while propagation and readout are recomputed; across the 17 node-intervention groups of the giant panda × BPA pairing, contributions decrease along ESR1 → GnRH → E2 (0.268, 0.198, 0.078), and the VTG node has zero effect because therian gating blocks this event (Supplementary Note 6). These interventions tie the attribution map to specific AON events, including the ESR1–GnRH–E2 sequence and the therian VTG gate.

The deepest two layers are propagation paths and knowledge sources. All intermediate quantities of a single prediction can be traced and recomputed, and the audit table for the BPA × giant panda pairing closes row by row over 582 rows, from input identity, gating three-state and receptor hits (ESR1 +0.741), through the five-step propagation path and the 53-dimensional feature vector (independent recomputation agrees with archived output to a maximum numerical difference of 1.1×10^{-16}), to GA²M per-feature contributions and the final probability (KE_OxStress p = 0.731); an audit table generated by the same procedure for the known AR agonist DHT has 497 rows. Every pathway relation retains AOP membership, KER identifier and evidence grade.

The audit combines exhaustive rule scans, 371 stratified case walkthroughs, official event checks, public sequence-database re-checks and known pharmacological-activity re-checks; literature-directed retrieval is assisted by an LLM agent and enters adjudication only after manual review. Nine rounds of adjudication produced 83 issue records, and each revision goes into an independent version whose impact scope is verified by new-versus-old field comparisons. The review triggered recomputation for all 756 species, and computable pairing–output-head records increased by 2,755,300 to 17,813,278, 33.7% of all 52.94 million pairing × output-head combinations; remaining cells are registered as gating abstentions rather than counted as negatives. The output of the known AR agonist DHT was thereby restored across all 756 species, controls

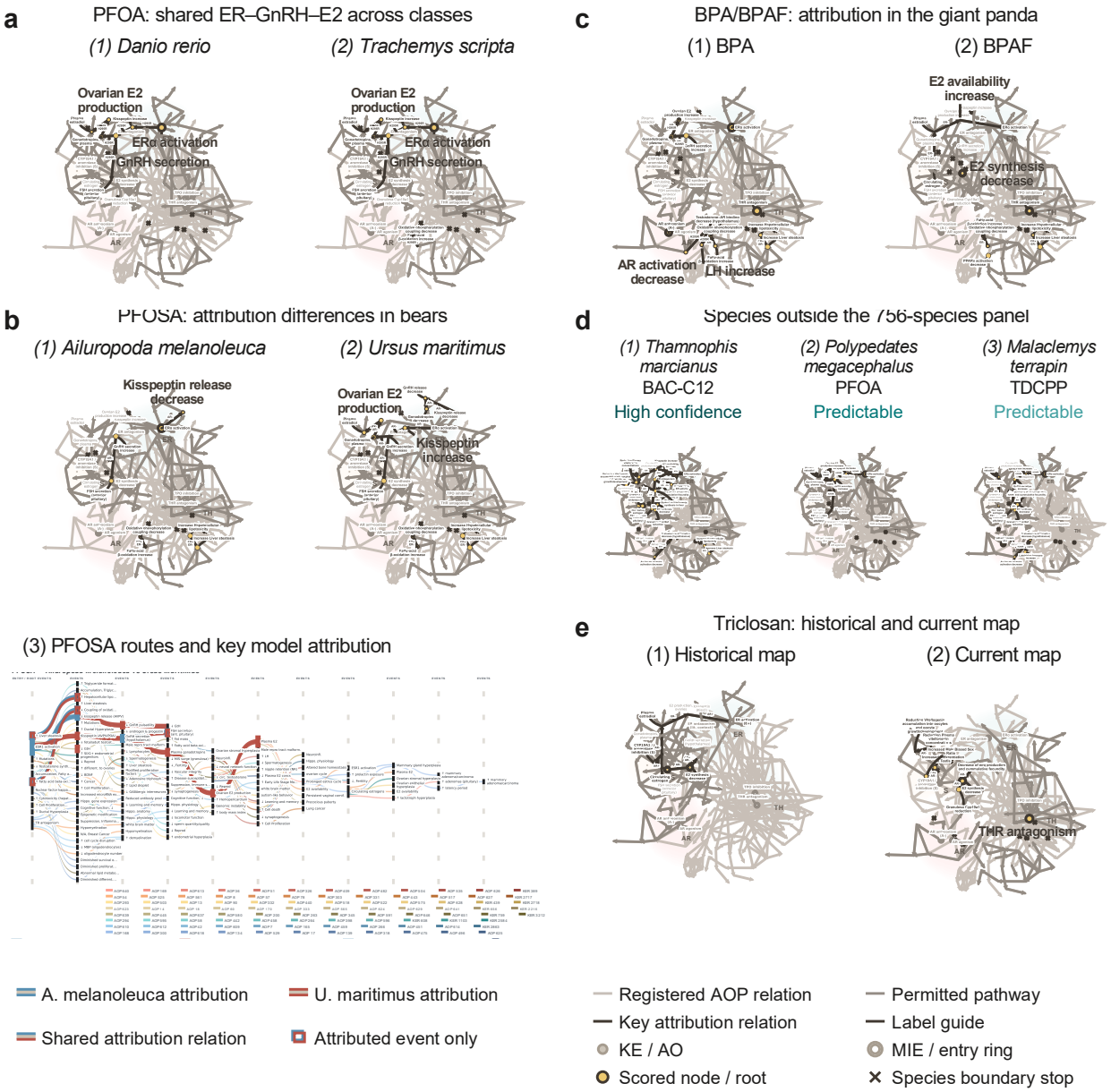
with genuinely no receptor basis still return abstentions, and the audit identifies both erroneous passes and erroneous rejections.

When model output conflicts with known evidence, revisions are routed by the stage at which the error sits. Evidence gaps or attribution problems are written back to the entry and relation registries; errors in permission, stopping or attribution anchors enter the rules and rebuild the affected routes; revisions that change training features rebuild the features and retrain as needed; official relation updates keep time versions and migration basis, so historical and current identities can be checked side by side. Conflicts that current evidence cannot adjudicate are registered as 30 validation targets with supporting or falsifying criteria, and subsequent experimental results feed back along the same provenance chain into the database and modelling specifications; one knowledge revision maps to a set of affected predictions, leaving a record of how new evidence changes existing judgements.

The case maps show these audit results and attributions for individual species–compound pairings. Ten species–compound combinations are drawn as 2D case cards around the E/A/S/T core pathways, presented in five groups and sharing the original hand-laid-out 200 nodes and 313 registered relations. Each map overlays the pairing's currently permitted pathways and the key model attribution of each qualified output head, and commonalities and differences among cases are compared on the same base map.

Zebrafish and red-eared slider × PFOA retain a common attribution backbone across distant taxa (Figure 8a). The two cards jointly highlight ERα, GnRH and E2-related events and share key attribution relations in the ER–kisspeptin–GnRH and gonadotropin–E2 regions. The functional link between kisspeptin and GnRH in zebrafish and the retention of the Kiss/KissR system in turtles provide the basis for this cross-class reproductive-regulation module[59,60].

682 **Figure 8**



683

684 Figure 8 | Commonalities, differences, evidence conditions and version audit of the endocrine-disruption mechanistic
685 maps. a(1)/(2), zebrafish/red-eared slider × PFOA, jointly highlighting ERα, GnRH and E2-related attribution. b(1)/(2),
686 giant panda/polar bear × PFOSA; b(3), Sankey expansion of the two cards' permitted relations, with bands corresponding
687 to AOP-Wiki registered chains and outlines distinguishing the two cards' key attribution and their shared part. c(1)/(2),
688 giant panda × BPA/BPAF. d(1)/(2)/(3), checkered garter snake × BAC-C12, spot-legged tree frog × PFOA, diamondback
689 terrapin × TDCPP. e(1)/(2), western tiger salamander × triclosan, the original historical map versus the current
690 mechanistic card. Ten independent pairings are drawn as eleven mini-maps, sharing the original 200 nodes, 313 relations
691 and hand-laid-out coordinates. Light grey is the registered background, medium grey the pairing's permitted pathways,
692 and dark grey the key attribution relations; gold cores mark entries projected from independent node or receptor
693 interventions, while endpoints of attribution relations not independently scored keep dark dots or entry rings without gold
694 cores. Double rings mark MIE or functional-entry references of the original base map, and crosses denote relations
695 stopped in the current view. Node size, gold cores and relation widths are drawn from normalised contributions of
696 qualified sources, with attribution of all output heads fully overlaid; full event names and item-by-item sources are given
697 in the standalone vector atlas.

698 Giant panda and polar bear × PFOSA present different attribution emphases on the same base map (Figure

699 8b). The panda card highlights the attribution relation from ERα to reduced kisspeptin release, whereas the

polar-bear card covers more attribution relations across the kisspeptin, GnRH, gonadotropin and E2 regions, with 4 and 10 key attribution relations in the two cards respectively. Species differences in the ligand response of mammalian ERα[61] provide an experimental basis for comparing species against the same receptor background, and FOSA has been detected in polar-bear brain tissue, giving this pairing a real exposure context[62]. The Sankey expansion overlays the two cards' attributions onto officially registered relations, so model judgements on close relatives can be compared node by node and relation by relation.

Giant panda × BPA/BPAF share similar pathway backgrounds but differ in key model attribution emphasis (Figure 8c). The BPA card highlights events and relations including reduced AR activation, reduced testosterone and AR binding, and increased LH, while the BPAF card highlights reduced E2 synthesis and relations around circulating-oestrogen–E2 availability. Enzymatic experiments showing direct inhibition of aromatase CYP19A1 by BPA and BPAF support the E2-synthesis-related attribution[63]. The two cards present 6 and 5 key attribution relations on the original base map, and the compound difference shows in attribution positions and relation combinations within the same mechanistic skeleton.

Checkered garter snake × BAC-C12, spot-legged tree frog × PFOA and diamondback terrapin × TDCPP sit in the high-confidence and predictable bands; all three original species lie outside the 756 precomputed-species catalogue, and after replay under the current rules their key model attribution remains stable (Figure 8d). The BAC-C12 card lists aromatase as an attribution node, consistent with experiments in which the C12 monomer inhibits aromatase and lowers cellular E2[64]; oestrogen-related transcriptional changes and fecundity decline after PFOA exposure in medaka[65] likewise support the reproductive-regulation attribution of that card.

The western tiger salamander × triclosan pairing shows the version update after mechanistic audit through its historical and current maps (Figure 8e). Early audits found the thyroid entry missing, and the corresponding pathway was restored after amphibian thyroid-related transcription and developmental evidence was supplemented[66]. Figure 8e(1) keeps the pre-revision map, and Figure 8e(2) shows the same species–compound card under the current rules, where key attribution involves THR, granulosa-cell Cyp19a1, E2 synthesis and reproduction-related events. Experiments showing that triclosan inhibits aromatase and lowers E2, together with in vivo evidence of disrupted gonadal differentiation in frogs, support the steroid-synthesis- and reproduction-related attribution of the current map[67,68].

Discussion

How can humans gain well-grounded knowledge of unknown life from limited observation? Can the shared evolutionary history of life turn limited experiments into knowledge that crosses species? The AOPedex we have built makes it possible to assess the endocrine-disrupting effects of zero-data species using receptor sequences or taxonomically assigned close-relative proxies. The model needs no toxicity records of the target species; receptor conservation is obtained by direct alignment of the species' own sequences and, where these are missing, approximated by k-NN from close relatives, with evidence gaps represented explicitly in the applicability-domain grades. The model outputs testable predictions that carry mechanistic pathways, confidence and applicability boundaries, and points to experimental entries for validating these predictions. In this framework, data estimate endpoint responses while mechanistic knowledge organises the basis of extrapolation; auditability lets that basis be checked and revised as new evidence arrives.

Interpretable models are often assumed to sacrifice discrimination[25]. In this work, the white-box EBM readout and the black-box TabPFN reach macro ROC-AUCs of 0.810 and 0.812 respectively, with no statistically significant difference detected, and on the sparsest endpoint, Morphology, the white box even overtakes slightly, under the same protocol used for the tabular foundation-model baseline[45]. The mechanistic increment is not a product of dimension inflation, since the equal-dimension control encoding produces no increment and shuffling the compound–feature correspondence abolishes the residual increment. Discrimination metrics, however, do not cover mechanistic direction. Randomising the polarity of every AON edge scrambles agonism and inhibition while leaving all other inputs untouched. The mechanistic output immediately distorts. Against the established mechanistic reference, the rank correlation of receptor-response ordering falls from 0.54 to below 0.03, direction agreement falls to chance level, and the macro AUC barely moves. Structure and activity strength are preserved, so discrimination does not flag the error; what breaks is direction, the quantity the mechanistic map conveys. Environmental machine learning has long treated discrimination metrics as the primary evidence of model quality, but such evidence does not reach the mechanistic layer. We therefore measure mechanistic fidelity independently, with direction agreement and signed rank correlation reported alongside AUC and R^2.

Post hoc explanation offers attribution tools for black-box models, but its conclusions remain subject to the method's own design. Attention weights require prediction-perturbation tests to establish their explanatory

value[69], SHAP values shift with the value function and reference data[70], and local surrogate explanations likewise rest on approximation assumptions[71]. Here, native white-box contributions retain an attribution-rank correlation of 0.987 after from-scratch retraining with changed seeds. Cross-model attribution rankings from Permutation SHAP for TabPFN and TreeSHAP for XGBoost correlate at 0.605 on the same task and feature set. Beyond feature contributions, AOPedex supports explicit in-model interventions at mechanistic nodes (Supplementary Table 23). In this white box, event identity, relation provenance, species permission and the native prediction readout are built directly into the computation chain. Attribution can be checked stage by stage, and anomalies located to entries, propagation relations, features or gating. Knowledge revision, rule updates or retraining are chosen by dependency scope and verified through versioned recomputation. Model interventions connect to toxicological experiments on the same mechanistic hypothesis. Methods such as Causal Shapley can analyse the direct and indirect contributions of interventions under a given causal structure[72]; the mechanistic map supplies this structure from the registered AON.

Cross-species extrapolation preserves compound ranking on the measured endpoints, and its cost concentrates at the input layer rather than in the readout. The trustworthiness of a species-level prediction is fixed before any output is produced, by the confidence of the receptor-conservation inputs, which the applicability-domain gating checks species by species. Errors in the mechanistic map can likewise be located and corrected. In the triclosan case, the mismatch between the pathway backbone and amphibian thyroid evidence was traced to the literature-registration layer, corrected, and replayed under the unified rules version by version (Figure 8e; Supplementary Note 6). Because registered knowledge can be corrected, the mechanistic map can be maintained as a scientific instrument.

Existing paradigms for cross-species extrapolation also include interspecies correlation estimation and species sensitivity distributions, the former predicting the sensitivity of untested species from the toxicity correlations of tested species and the latter extrapolating protective thresholds from multi-species toxicity distributions; both presuppose in vivo records for the target taxon, and quantitative read-across solves the analogous problem on the compound side. For species lacking even these records, we output traceable event chains and entry evidence, which can be cross-validated against existing paradigms on species with shared data. Against ICE interspecies regressions, ranking remains comparable (median Spearman difference +0.046, 95% CI crossing zero), while the concentration-calibration MAE difference is +0.088 log10. Concentration outputs are interpreted in a screening setting unless concentration–response data are available for calibrated use

(Supplementary Table 17). AOPedex therefore complements ICE and related approaches by pairing comparable species ranking with traceable event chains.

The white box also has costs. Point prediction on the continuous endpoints still trails the strongest black box (macro ΔR^2 −0.0241), and the cost of the interpretable readout currently concentrates on the continuous side. The transparency of the EBM comes from its functional form of main effects plus limited second-order interactions, a form that cannot express higher-order interactions or coupling with external factors; on the ER-agonism endpoint, interaction gains concentrate in the high-activity tail while the mid-activity band shows slight negative transfer, so the additive approximation does not hold everywhere. These results identify richer interaction terms as the next modelling step for continuous endpoints.

The Morphology reversal mentioned above comes from a single endpoint under a single splitting protocol and should not be over-read. Prospective external sets and systematic sample-size sweeps can test whether this behaviour extends beyond the present endpoint and split.

This work has clear boundaries, and each points to a concrete extension. First, the mechanistic space is limited to vertebrate E/A/S/T. The HPG and HPT axes are highly conserved across vertebrates, whereas invertebrates possess taxon-specific endocrine pathways whose mechanistic evidence remains incomplete; as that evidence accumulates[54], the same framework can extend the event network and entry registration taxon by taxon. Second, thyroid supervision currently rests on the binary hTPO head, and extending event registration along the cross-species thyroid AOP network[34] together with selectively adding thyroid-axis concentration–response data is the direct prerequisite for restoring continuous modelling. Third, the mechanistic map outputs qualitative paths carrying direction and support levels, not quantitative risk values; concentration–response relationships and in vivo exposure conversion require measured toxicokinetic parameters, and introducing TK/ADME models under mechanistic-prior constraints is the key step in upgrading qualitative paths to quantitative predictions[13,73], with quantitative assessment also needing checks on how species coverage and acute and chronic endpoints affect uncertainty[74]. Mechanism-weighted ecological risk assessment has already been pursued at the China scale[75], and its exposure-weighted route is complementary to the mechanistic-susceptibility route here, the two able to converge on the same species–compound grid. Fourth, the current assessment is static and excludes exposure scenarios, individual traits such as age and body size, and mixed exposures; the next step is to attach exposure scenarios, life stages and mixture-component interactions to the same white-box skeleton[73,76], with leave-species-out extrapolation and prospective experiments testing each

addition. Which mechanisms can be inherited, and under which biological conditions does this inheritance fail? From molecular entry to downstream outcome, which step is most likely to differ among species? These questions await the completion of the research agenda above, and filling in the white-box architecture's coverage of multi-source fusion and cross-validation will in turn broaden the counterfactual queries it can answer, bringing it a step closer to a 'world model'[77–79].

Data quality remains the fundamental constraint of machine learning in the environmental sciences. With publication volumes growing rapidly, the data quality of much published work is insufficient to support its conclusions, and model output is not mechanism[47,80,81]; the polarity-randomisation ablation illustrates this, with discrimination scores unmoved while mechanistic ordering has dropped to chance level. A white box exposes the inside of a model but does not remove the spurious associations that data bring; removing them takes domain-wide coordination on data production and sharing[13] together with community norms for leakage control and baseline reporting[82]. A viable route is to organise knowledge and locate attribution in an auditable white-box framework, reserving complex models for the critical positions where evidence is sufficient. LLM agents have begun to take on auditable automation in environmental modelling[37,83], while hierarchical mechanistic modelling[84] and causally constrained graph learning[36] address complementary parts of the same problem. This audit loop implements the corresponding human–machine division of labour in toxicological knowledge maintenance. Within this workflow, AOPedex keeps predictions, mechanistic assignments and evidence updates in one auditable record.

Why have white-box extrapolation models remained so rare? The usual answer cites performance and architecture[25]. A less visible obstacle is keeping the mechanistic evidence current[85]. A mechanistic prior encodes the evidence as it stood at build time, and evidence keeps changing. By the time the model is finished, new studies may already have overturned its conclusions. Extrapolation enters precisely the region where evidence is thinnest and conclusions date fastest. Historically, auditing mechanistic evidence against new studies was unaffordable: literature outgrew human curation, the information gap widened yearly, and mechanistic priors were expected to go stale. With a black box, the builder bears no such obligation, because the mechanistic commitments are never made explicit and decay goes undetectable[86]. We therefore pair the mechanistic prior with an evidence audit that checks model outputs against new findings and propagates adjudicated revisions. Versioned evidence audit makes this decay visible, localises the affected commitments and routes revisions back into extrapolation.

AI assistance changes this cost structure[37]. With literature monitoring, candidate retrieval and evidence extraction running as AI-assisted, human-adjudicated workflows[87], the audit that once demanded a team now demands a pipeline[88]. The white box suits this workflow because its explicit commitments let every finding map onto a localised revision, and its simple functional form keeps every commitment individually checkable. Mechanistic prior, model output, evidence audit and iterated update form a cycle whose cost falls as AI tools spread; adjudicated evidence updates the affected entries and predictions.

For toxicological experiments, the value of such models lies in directing resources to specific mechanistic tests. Data vacuum, data heterogeneity and data bias leave many existing records unable to support training, and more data of the same kind will not fix this. The model maps validation needs to specific AOP key events and identifies the mechanistic steps most lacking experimental evidence, to which experimental resources can flow; the new data feed back into modelling, and more accurate models guide the next round of experimental design[89]. For threatened species, this capability helps identify endocrine-disrupting effects that could impair survival and reproduction, providing a basis for protecting populations already at risk. The evidence audit checks the model against the literature, and experiments audit the model against the bench. The two loops run in parallel, and each supplies the other with its next task.

Methods

Data, identity keys and supervised endpoints

Mechanistic knowledge is taken from AOP-Wiki, compound activities from the dose–response fittings and assay annotations of EPA ToxCast invitrodb v4.3, and receptor-pocket sequences from orthologous protein records in UniProt, NCBI and Ensembl. Compounds are keyed across data sources by DTXSID. The unified catalogue contains 9,458 compounds, of which 8,806 have a complete 2,058-dimensional structural core and 53-dimensional mechanistic view. The 12-receptor activity spectra cover 5,387 compounds and 19,557 compound–receptor pairs, defining the compound-side injection scope of the AON. Untested assays keep their untested flag and detection negatives are recorded under the original detection rules; the two are never interchanged, and combinations without a registrable entry are registered as map-gating abstentions. The species-side pocket pool contains 1,118 chordate species; the applicability-domain analysis and case maps in the main text use the 756 species (751 genera) selected by genus-level representation and sequence-quality rules.

Supervision tasks are registered by original measurement type. The endocrine axis comprises five continuous endpoints (ER agonism, ER antagonism, AR agonism, AR antagonism and H295R steroidogenesis) and one binary endpoint (hTPO inhibition), and the general-toxicity axis comprises three human KEs and four zebrafish-embryo AO binary endpoints. Compounds with at least one label number 1,858 for the five endocrine continuous tasks and 5,248 for the general-toxicity endpoints, with 1,604 shared by both axes; the binary hTPO endpoint has 288 observations. Each head consumes only compounds labelled for its own endpoint, and missing labels are not treated as negatives. The compound-to-mechanism correspondence also defines the coverage boundary. The independent exploratory analysis of continuous TPO potency had 147 valid labels and was negative in all 15 confirmation cells, failing the confirmation criteria of the main task. Thyroid supervision therefore uses the binary hTPO head in the main text, while mechanistic structure provides quantitative correction on endocrine axes already covered by qualified paths. ECOTOX records serve species holdout and record-level direction checks; EPA EDSP Tier 1 records provide an independent overlay for dual-axis candidates that never enters model fitting.

The AOP event network and relation eligibility

One hundred and seventy-three endocrine-related AOPs were screened from AOP-Wiki and merged into a directed subnetwork of 629 nodes spanning MIEs, KEs and AOs. The relation table contains 711 provenance-traced edges, each recording AOP membership, relation identifier, adjacency and evidence. Of these, 541 qualify under the adjacent-KER and AOP-criteria rules and enter stepwise propagation, while non-adjacent summary relations are retained for knowledge display and not used as one-step propagation edges. Node polarity is read from event semantics and encoded as +1, −1 or 0 by the event-direction lexicon, and nodes without a definite polarity carry no signed propagation.

The twelve official receptor or enzyme entries are ESR1, AR, GR, AHR, CYP19A1, THR, PPARG, PPARA, PXR, CAR, MR and RAR. Receptor injection sites are registered only from events carrying the MIE flag; the receptor label of a whole AOP is not extended to its events, and generic stress events are not treated as receptor MIEs. PGR lacks a qualified MIE in the current subnetwork and therefore does not enter the official entries. Taxon rules are fixed before propagation. VTG target nodes are switched off for therians, whereas monotremes are not subject to this gate. Network relations, the direction lexicon and the entry registry each keep version and provenance records.

Receptor-pocket conservation and species inputs

The ligand-binding pocket sites of each receptor are defined by structural templates; for example, the 1ERE template for ESR1 contains 16 sites, the 2AM9 template for AR 21 sites, and the CYP19A1 template 37 sites. After orthologous sequences are aligned to the human reference, pocket conservation is calculated over valid aligned template positions:

$$C_{s,r} = \frac{1}{|\mathcal{P}_{s,r}|} \sum_{p \in \mathcal{P}_{s,r}} \mathbf{1}\left(a_{s,r,p} = a_{\mathrm{human},r,p}\right).$$

Here, $a_{s,r,p}$ is the residue aligned to human-reference position p, and $\mathcal{P}_{s,r}$ contains registered pocket positions with non-gap residues in both sequences. Profiles failing the registered coverage or receptor-identity checks remain missing. Anomalous isoforms, receptor-identity mismatches and low-quality pocket sites are flagged, and missing entries stay missing rather than being back-filled from older versions. The choice of binding pocket over full-length sequence as the primary input is consistent with the key-residue basis of SeqAPASS[27,41].

Each species–receptor value records its source as one of five types — the species' own sequence, a genus-level proxy, a neighbour estimate, a global-median fallback or a default. When the target species lacks its own receptor sequence, a sourced genus-level value is used first, and the remaining gaps are filled by cross-genus 5-NN from sequenced close relatives or by the global median, according to the phylogenetic signal of the receptor. By Mantel r ($r \geq 0.4$ for the strong-signal layer and $r < 0.15$ for the weak-signal layer), PPARG, CYP19A1, PPARA, PXR and GR form the strong-signal layer, AR and AHR the mid-signal layer, and MR, CAR, ESR1, THR and RAR the weak-signal layer; the first two layers allow neighbour extrapolation, while the weak-signal layer uses the global-median fallback. The default conservation where no value is available is 0.8, and its default-fill provenance is reported with it. Conservation source, reference distance and neighbour dispersion therefore grade the reliability of the mechanistic point estimate.

Deterministic propagation and mechanistic representation

ToxCast activities are called at hitc ≥ 0.9 and summarised per receptor across assays by activity direction, AC50 and response magnitude normalised within each testing platform, yielding the signed receptor spectrum $\mathbf{S}_c = (S_{c,r})_{r \in \mathcal{R}}$. Where assays conflict in direction for the same receptor, the conflict flag is retained rather

than resolving biological effect from a single screen hit. For compound c, species s and receptor r, the injection value is

$$\begin{aligned} \tilde{I}_{c,s,r} &= S_{c,r}\,\hat{C}_{s,r}\,W_c, \\ I_{c,s,r} &= \tilde{I}_{c,s,r}\,\mathbf{1}\big(|S_{c,r}| \geq 0.01\big)\,\mathbf{1}\big(|\tilde{I}_{c,s,r}| > 0.01\big). \end{aligned}$$

The quantity $\hat{C}_{s,r}$ is the own-sequence conservation value or the registered fallback estimate. When the trigger-evidence layer is enabled and the compound has a registered tier, direct chemical triggers, receptor-axis inference and the registered no-trigger tier receive $W_c = 1.0$, 0.8 and 0.5, respectively. Task-homologue-masked endocrine views that do not enable this layer use $W_c = 1$; the engine also retains $W_c = 1$ when no tier lookup is available. Trigger tiers affect injection magnitude, not the provenance of species conservation or the predictability state. Receptors failing either injection threshold receive no injection.

After injection is mapped to the corresponding MIEs, the signal propagates five steps along the 541 qualified edges. Each step preserves existing activations and normalises the outgoing edges of each upstream node by out-degree; the edge direction is the product of the upstream and downstream event polarities, and node values are clipped to [−10, 10]. The edge weights and initial node state are

$$w_{uv} = \frac{p_u p_v}{d_u^+}, \qquad \delta_{c,s,v}^{(0)} = \sum_{r \in \mathcal{R}} L_{v,r}\, I_{c,s,r}.$$

Here, $p_v \in \{-1,0,1\}$ is the registered event polarity, d_u^+ is the out-degree in the qualified propagation graph and $L_{v,r}$ is the registered signed receptor-to-MIE entry coefficient. The update is

$$\delta_{c,s,v}^{(t+1)} = \mathrm{clip}_{[-10,10]}\left[\delta_{c,s,v}^{(t)} + G_s(v) \sum_{u:(u,v)\in\mathcal{E}_{\mathrm{prop}}} w_{uv}\,\delta_{c,s,u}^{(t)}\right], \qquad t = 0, \ldots, 4.$$

Here, $G_s(v) = 0$ for a species-gated target node and 1 otherwise; the clip operator is applied after every step. Target nodes whose species gate is closed receive no propagation. Human and zebrafish serve as the human-source and fish-source references for the supervised models; for any target species, the same network and propagation rules are applied once that species's conservation values are entered.

Final node activations are compressed into the 17-dimensional AON representation as signed means over event groups — four endpoint tendencies, ten key-KE biological values and three KE-anchor tendencies. Duplicate aliases of the oxidative-stress event group are merged into one feature:

948
$$z_{c,s,g} = \frac{1}{|\mathcal{V}_g|} \sum_{v \in \mathcal{V}_g} p_v \, \delta^{(5)}_{c,s,v}, \qquad g = 1, \ldots, 17,$$
$$\mathbf{x}_{c,s} = \left[\mathbf{S}_c, \mathbf{P}_c, \mathbf{E}_c, \mathbf{z}_{c,s}\right] \in \mathbb{R}^{53}.$$

949 The set $\mathcal{V}_g$ contains the registered nodes of event group g with non-zero polarity; its mean is defined as zero
950 when the set is empty. The vectors $\mathbf{S}_c$, $\mathbf{P}_c$ and $\mathbf{E}_c$ are the 12-dimensional chemical-side signed-activity,
951 potency and efficacy channels, and $\mathbf{z}_{c,s}$ is AON17. Head-specific views $\mathbf{x}^{(h)}_{c,s}$ are materialised after task-
952 homologue masking and re-propagation. The structural core comprises the 2,048 bits of ECFP4 and ten
953 physicochemical descriptors. The AON computation fits no parameters, and the supervised heads learn
954 readout functions only on these two representations.

955 **Readouts, unified splitting and dual-axis convergence**

956 The thirteen supervised heads share the DTXSID catalogue and the outer structural split, with losses separated
957 by endpoint measurement type. The unified split builds 3,894 scaffold groups over the 8,806 modelable
958 compounds, grouping by canonical Bemis–Murcko scaffold and further subdividing large groups and acyclic
959 compounds by ECFP4/Butina clustering; five fixed seeds each yield five folds, and compounds of the same
960 group never cross folds. General-toxicity classification uses all five seeds, and the endocrine continuous heads
961 use four of them as confirmation seeds. Preprocessing, imputation and calibrators are fitted on the training
962 folds only. On the general-toxicity classification side, LR, XGB, TabPFN and EBM are compared, with the
963 53-dimensional mechanistic view as the primary input. The primary hTPO comparison uses the structural
964 feature arm on the same outer folds. The thyroid-lineage-masked mechanistic view is a separate control. The
965 black-box control is TabPFN, a strong black-box baseline for small-sample tabular data[45], and it requires no
966 tuning. Classification performance is reported as ROC-AUC, PR-AUC, Brier score and ECE, and continuous
967 performance as R^2, Spearman, RMSE, MAE and calibration slope. Tuned gradient-boosting and tuned MLP
968 controls (32- and 20-configuration random searches, hyperparameters selected on inner three-fold splits)
969 guard against bias from untuned baselines, and a structure-only arm, gradient boosting on the 2,048-bit
970 ECFP4 fingerprint, locates the signal source in the feature composition; per-endpoint metrics and paired
971 intervals for both control sets are in Supplementary Tables S3 and S4.

972 The structural base score of the five endocrine continuous heads is given by ridge regression on the 2,058-
973 dimensional structural features. For each outer training fold, the ridge base score is first cross-fitted within

that training fold, and a 53-dimensional mechanistic EBM is then trained on the residuals of the training samples; the final prediction on the outer test fold is the sum of the ridge structural prediction and the mechanistic residual prediction:

$$\begin{aligned} r_{i,h}^{(-k)} &= y_{i,h} - \hat{b}_h^{(-k,-\ell(i))}(\mathbf{u}_i), \\ \hat{y}_{c,s,h}^{(-k)} &= \hat{b}_h^{(-k)}(\mathbf{u}_c) + F_h^{(-k)}\left(\mathbf{x}_{c,s}^{(h)}\right). \end{aligned}$$

Here, $\mathbf{u}_c \in \mathbb{R}^{2058}$ is the structural core, k is the outer test fold, and $\ell(i)$ is the inner fold containing training compound i. The base predictor $\hat{b}_h^{(-k,-\ell(i))}$ excludes both folds; $F_h^{(-k)}$ is fitted on these cross-fitted residuals within the outer training set. Residual ridge, XGB and TabPFN share the same structural base score, residual labels and mechanistic input, and the pure structural EBM-S provides the nonlinear structural control. The ER, AR and H295R mechanistic views, together with the thyroid mechanistic control, are masked on assay-derived values homologous to the task's labels, and AON17 is re-propagated after masking the direct columns and aggregated values, preventing label provenance from flowing back through receptor channels or propagation nodes. General-toxicity probabilities are output as independent heads; cross-head late fusion and endocrine probability bridging are examined on out-of-fold predictions only. The masked views are verified directly. The three views match the contract on column order and counts; masked columns are all zero across the 9,458 compounds; and no residual feature is a deterministic copy of a label (Supplementary Section S1.2).

Dual-axis candidates are classified by the out-of-fold general-toxicity probability and the endocrine response of the same compound. Convergence takes per-compound means of confirmation-seed OOF predictions; the general-toxicity coordinate is the equal-weight mean over available classification endpoints, and the endocrine coordinate converts each continuous endpoint's predicted potency to within-endpoint percentile ranks before equal-weight averaging over available endpoints. A pre-specified probability threshold of 0.5 and the cohort median of 0.5 divide the supervised compounds into four candidate regions; the 32 pre-specified EPA EDSP Tier 1 compounds serve only as an independent overlay, with correspondence described by an odds ratio and 95% CI, and external labels never enter model training or re-set quadrant thresholds.

Readout direction is validated against a pre-specified control panel. The panel contains 20 positive controls and 10 abstaining negative controls, each carrying both in-library labels and external authority, with direction agreement, entry-receptor hits and per-axis cleanliness counted separately (item-by-item results in Supplementary Note 2 and Supplementary Table 9).

Representation ablation and the ER-agonism white box

Representation ablation on the classification axis compares 29D (12-dimensional signed direction + AON17) with 53D (adding the 12-dimensional potency and 12-dimensional efficacy channels), with EBM and TabPFN each retrained from scratch. A supplementary 2 × 2 factorial design compares these two cells under fixed outer-split seeds together with 41D, which splits direction into separate positive and negative channels (AON17 + 24-dimensional dual-channel direction), and 65D (41D + 12-dimensional potency + 12-dimensional efficacy). On the endocrine axis, the mechanistic increment is measured by the paired difference of the residual EBM against the pure structural EBM-S, and a joint permutation of the mechanistic features tests the compound–mechanism-input correspondence. For the structural white box of ER agonists, the purely additive E0 is compared with E8, which allows interactions; both are compared on confirmation out-of-fold predictions, bits 1313/1778/0745 of the ECFP4 fingerprint are decoded to their actual atomic environments, and native main-effect functions, bit × logP interaction functions and single-compound summed contributions are read off. For the exploratory direction check, 48 pairs of phenol-ring structural counterparts selected under a fixed single-bond-cut rule are evaluated under a shared-compound design (Figure 4).

Nested direct-Ridge comparisons used the 2,058 structural features with or without task-masked receptor channels. Four split seeds and five outer folds were retained, and four inner folds selected regularization from six values separately for each arm. Nonbinary variables were imputed and standardized using training data only. Metrics were averaged across seeds and then tasks, with 5,000 paired scaffold-bootstrap resamples.

Mechanistic-fidelity gauges and controlled perturbations

The AON mechanistic ablation compares, under the same readout and outer folds, the true network against no propagation, stratified-degree polarity-preserving rewiring, edge-polarity randomisation, node-identity permutation, removal of the species gate and six-step propagation. Literature direction anchors cover 59 pairs under the event-direction lexicon (62 rows in the source table); direction agreement and signed rank correlation measure mechanistic semantics, ROC-AUC measures discriminative response, and the two metric families are aggregated separately. The receptor-input, topology and literature-registration layers each carry independent audit records, and the model's endpoint predictions do not serve as experimental support for pathway edges. Independent direction anchors are taken from EPA EDSP Tier 1 weight-of-evidence conclusions (19 pairs) and primary literature (1 pair), excluding AOP-Wiki sources, with positive and

negative agreement reported separately and disagreements decomposed into definition differences, threshold-boundary cases and genuine conflicts.

Cross-species transfer evaluation

On the four zebrafish AO endpoints, readouts are trained under the same compound folds and measured labels on human-source AON17, fish-source AON17 and the raw receptor-injection spectrum; the reference-species parameterisation difference is reported per endpoint and macro-averaged over the four endpoints:

$$\Delta_h^{\mathrm{ref}} = \mathrm{AUC}_h^{\mathrm{fish}} - \mathrm{AUC}_h^{\mathrm{human}}, \qquad \overline{\Delta}^{\mathrm{ref}} = \frac{1}{4} \sum_{h \in \mathcal{H}_{\mathrm{ZF}}} \Delta_h^{\mathrm{ref}}.$$

Here, AUC denotes ROC-AUC and $\mathcal{H}_{\mathrm{ZF}}$ contains the four zebrafish AO endpoints; positive Δ_h^{ref} denotes higher discrimination under the fish reference. Sequence and pocket errors of mammal-to-fish extrapolation are benchmarked against within-class holdout of same-class fish on the same receptor set. The per-compound transfer difference compares the two parameterisations on predictions of the five endocrine endpoints, reporting the median and p90 of |Δ| in units of the label standard deviation, the direction-flip rate at zero crossing and the pooled Spearman correlation; the flagship subset covers 65 registered cases × 5 endpoints × 5 seeds (1,625 cells), the library-scale arm re-scores all 9,458 compounds under both references, and a no-propagation arm separates direct receptor injection from network amplification. Reference-species robustness is examined by swapping medaka or fathead minnow for zebrafish. The design draws on CORDIAL[52] and Interformer[90].

Mechanistic resolution is assessed under unified receptor-conservation inputs at six phylogenetic-distance tiers (16 species pairs per tier). Each pair generates AON17 with the same 24 receptor probes, and mechanistic-vector distance is compared with 12-receptor input distance; receptor-level within-class and cross-class differences, and the receptor-dominant compound subsets of the two close-relative fish pairs, identify the sources of local resolution. Two further checks — the correlation of own-sequence provenance with distance, and a 2 × 2 ablation of continuous conservation versus VTG gating — address whether the separations in Figure 5e–g come from data provenance or from the taxon gate alone.

Extrapolation fidelity uses one fixed set of 220 compounds that each activate at least two official receptors. A compound–species pair enters the mechanistic-vector reference set only if all active receptors of the target species have pocket conservation from the species' own sequence, giving 65,457 pairs across 904 species, 639

genera and five classes. LOSO, LOGO and LOCO hold out species, genera and classes respectively, with the neighbour pool, the receptor-stratified fallback and the global medians rebuilt in each fold; the stratified rule is compared against uniform 5-NN and global-median baselines, and the primary error between the 17-dimensional estimated vector and the own-sequence reference vector is

$$D_{c,s} = \frac{\|\hat{\mathbf{z}}_{c,s} - \mathbf{z}_{c,s}\|_2}{\|\hat{\mathbf{z}}_{c,s}\|_2 + \|\mathbf{z}_{c,s}\|_2 + \varepsilon}, \qquad \varepsilon = 10^{-12}.$$

Here, $\hat{\mathbf{z}}_{c,s}$ uses fold-rebuilt conservation estimates and $\mathbf{z}_{c,s}$ uses the target species's own-sequence pocket values. Distances aggregate over compound, species, genus and class and are macro-averaged with equal class weights; the 95th percentile of per-species compound errors defines the severe-error tail.

Sensitivity to paralogue choice replaces the originally selected sequence with the next-best candidate-pool sequence (397 cells across 319 species), annotating ESR1-pool entries as esr1/esr2 and recomputing the three states of all 756 species (Supplementary Table 14). Record-level checks use ECOTOX compound–species–endpoint records. The exports do not include hormone-change direction, precluding agreement analyses for that field. Record-level direction is instead defined as ESR1 binding strength multiplied by target-species conservation, with AR binding as the negative control; exposure and label permutations test the direction and specificity of the association. Species-holdout evaluation uses GroupKFold by species, comparing prediction losses under close- versus distant-relative conservation replacement, with sensitivity to unmeasured confounding expressed as an E-value[91]. All three ECOTOX exports are therefore used for qualitative-consistency and species-holdout discrimination analyses (quality-gate chain and label-agreement rates in Supplementary Table 17); the ICE-style control evaluates ranking and concentration calibration on 182 species pairs of log–log regression across 25 assessable target species (Supplementary Table 17).

Here, $\beta_{0,h,b}$ is the intercept, $f_{j,h,b}$ is a main-effect function and $\mathcal{J}_{h,b}$ is the set of fitted feature pairs for head h and seed b. Continuous endocrine heads add the structural ridge base $B_{c,h,b}$, whereas binary heads apply the logistic link:

Applicability-domain gating and threatened-species matching

The applicability-domain state is determined by the minimum conservativeness of the four core receptors (ESR1, AR, THR and CYP19A1) together with endpoint applicability. A species is unpredictable if either the core minimum conservativeness or the maximal endpoint applicability falls below 0.05; among the remaining

species, those with core minimum conservativeness below 0.4 are predictable, and those meeting the thresholds are high-confidence. Close-relative evidence is provided by the phylogenetic-distance matrix, with species divergence times taken from TimeTree[92] and species-level neighbour counts and receptor-level 5-NN recorded separately. State summaries use the 756-species genus-representative panel, and IUCN VU, EN and CR statuses are matched to that panel by identity only, with species absence never read as low risk. Sensitivity of the gating parameters is recomputed tier by tier; perturbations of the hitc threshold, propagation steps, conservativeness weight and default fill change no species' three-state assignment, while tightening or relaxing the three-state gating thresholds by one tier (about 20% relative change) moves coverage to 635 (84.0%) and 742 (98.1%) species respectively (Supplementary Table 20).

Mechanistic-map generation, intervention and evidence audit

The interactive 3D mechanistic map takes the species–compound pairing as its retrieval unit and presents endocrine-disruption pathways and key model attribution with their sources. Knowledge display and numerical propagation each fix their relation sources: the web relation set contains 716 relations checked against an independent official KER registry export, plus 3 historical official BDNF relations that keep their original identifiers and endpoints; relation admission follows the official KER records for event identity, adjacency and direction, and the review status of historical migrations is retained. Numerical propagation runs along the 541 propagation-eligible edges.

Attribution holds the pairing, output head and model version fixed and recomputes AON17 and the task-specific EBM readout after one of three interventions: zeroing a single node's signal at every propagation step, removing a single propagation relation, or switching off one receptor injection together with its receptor features. For heads using the 53-dimensional mechanistic view, the native EBM score is

$$F_{h,b}(\mathbf{x}) = \beta_{0,h,b} + \sum_{j=1}^{53} f_{j,h,b}\,(x_j) + \sum_{(j,\ell)\in\mathcal{J}_{h,b}} f_{j\ell,h,b}\,(x_j, x_\ell).$$

Here, $\beta 0, h, b$ is the intercept, fj, h, b is a main-effect function and $\mathcal{J}h, b$ is the set of fitted feature pairs for head h and seed b. Continuous endocrine heads add the structural ridge base Bc, h, b, whereas binary heads apply the logistic link:

$$\hat{y}_{c,s,h,b} = B_{c,h,b} + F_{h,b}\left(\mathbf{x}_{c,s}^{(h)}\right), \qquad \hat{p}_{c,s,h,b} = \frac{1}{1+\exp[-F_{h,b}(\mathbf{x}_{c,s}^{(h)})]}.$$

The first expression defines a continuous response and the second a binary probability; they refer to the respective fitted heads. For a continuous head, the structural ridge base stays fixed under pairing-fixed interventions. For intervention a, the signed contribution difference is

$$\begin{aligned} \Delta_{a;c,s,h,b} &= \hat{y}_{c,s,h,b} - \hat{y}^{(a)}_{c,s,h,b} \\ &= F_{h,b}\left(\mathbf{x}^{(h)}_{c,s}\right) - F_{h,b}\left(\mathbf{x}^{(h,a)}_{c,s}\right). \end{aligned}$$

The view $\mathbf{x}^{(h,a)}_{c,s}$ is rebuilt after the node, edge or receptor intervention, with task masking and the corresponding receptor-feature changes retained. Labelled compounds use their held-out-fold models, unlabelled compounds five-fold ensembles. Within a pairing, output head and intervention type, the mean absolute difference is normalised over qualified members:

$$\overline{d}_a = \frac{1}{5}\sum_{b=1}^{5} |\Delta_{a;c,s,h,b}|, \qquad N_a = \frac{\overline{d}_a}{\max\limits_{q \in \mathcal{A}^{\mathrm{type}(a)}_{c,s,h}} \overline{d}_q}.$$

The set $\mathcal{A}^{\mathrm{type}(a)}_{c,s,h}$ contains members with a valid pathway anchor, in-range intervention features and a passing observation-state gate. If the maximum is zero, N_a is defined as zero. A member is displayed when $N_a \geq 0.35$, it ranks in the top ten of qualified non-zero members in at least four of five seeds, and its direction agreement reaches 0.8. All 1,182 intervention scenarios per head (629 nodes, 541 propagation relations, 12 receptors) are enumerated exhaustively. Members passing these gates are merged, projected by source and may form multiple segments.

The scientific audit fixes the inputs and the model baseline and then runs rule checks, stratified case reconstruction, official-event and sequence re-checks, pharmacological-consensus re-checks and issue adjudication. Directed literature retrieval runs as a two-stage human–machine workflow. For each anchor claim, an LLM agent proposes candidate references with supporting evidence excerpts[37,83], and only candidates that survive full-text manual review enter the adjudication issue table; the agent is confined to retrieval and excerpt preparation, and adjudication, revision and registration remain with human curators. Approved items enter a new version rather than overwriting the baseline, and repairs are verified field by field against the approved change list, with full recomputation against fixed baseline expectations and species-level negative controls. Experimental candidates pre-register hypotheses, readouts and supporting or falsifying criteria, and measured results re-enter the workflow as new evidence.

The ten 2D case cards of Figure 8 use a common endocrine base network (200 nodes, 313 registered relations) at fixed scale, with edge-by-edge AOP provenance. Pathways show only candidate relations that carry an entry trigger and are permitted by the target species' branches, with evidence grades marked separately; key model attribution is projected from the task-specific EBM's feature contributions onto events and displayed at a fixed threshold of 0.35 after feature–head consistency and cross-fold stability checks. Signed propagation support and species-switched-off branches carry their own symbols and are never merged with literature relations or EBM contributions into a single causal edge. The four original species outside the precomputed catalogue are replayed under the same engine, models and gates. Element-level PMIDs and the SUPPORTED, BY_ANALOGY, CONTRADICTED and NO_EVIDENCE adjudication tables are in Supplementary Note 6 and Supplementary Data 6.

Statistics, versioning and reproducibility

Model comparisons on the classification axis and the 53D–29D differences use sample-level paired bootstrap over compounds (2,000 resamples) on out-of-fold probabilities pooled per compound across the five seeds; the endocrine continuous axis uses scaffold-cluster paired bootstrap (5,000 resamples) on per-compound mean predictions of the four confirmation seeds. Comparisons of the endocrine residual head against the structural head, and of the white box against the black box, use paired resampling over the same outer scaffold groups; species-extrapolation errors use a four-tier bootstrap over class → genus → species → compound (2,000 resamples), with the severe-error tail resampled along the same hierarchy. Macro means weight endpoints or the five classes equally, and the unit is annotated in each table or figure. Calibration and structural interpretation both use out-of-fold predictions; training-set attribution is not generalisable interpretation. Permutation SHAP comparisons use 300 samples and 50 background sets. Figure 2a, c, d are computed from the seed-mean predictions of the five classification seeds, and Figure 2b from the four confirmation seeds of the endocrine continuous heads.

Reproduction inputs include the unified compound-structure and receptor-activity views, the DTXSID/scaffold outer split, the event-direction lexicon, the 541 propagation-eligible edges and the task-masked views. Each analysis registers the models, inputs and source tables used, and reproduction files are checked against file checksums and output denominators. The core scripts, finalized derived inputs and

redistributable starting data are released together; source identifiers and available third-party versions are recorded in the data inventory.

Source-workbook labels were reconstructed using task-specific transformations, and training and test sets were compared across 125 task–seed–fold partitions. Graph features were recomputed from task-masked receptor inputs. Label and graph-feature agreement was within 10^{-6} (Supplementary Table 2).

Author contributions

Y.Q.Z. implemented the computational framework, developed the methodology and software, curated and analysed the data, trained and validated the models, performed the ablation, robustness and cross-species extrapolation analyses, generated the results and visualisations, built the interactive web resource and reproducibility assets, and wrote the original draft and subsequent revisions. G.H.C. developed the AI-assisted literature workflow, including literature retrieval and evidence-excerpt preparation, manually reviewed and adjudicated the extracted evidence, and optimised the scientific figures and visual presentation. N.L. constructed and curated the intermediate molecular toxicity pathway network, organised the key events and their relationships, and checked network topology, directionality and evidence provenance. Y.M.J. reviewed and validated the upstream chemical–target interaction mechanisms, checked receptor and molecular-initiating-event assignments, and assessed the mechanistic plausibility and direction of chemical–protein interactions. Y.G.G. evaluated the downstream adverse outcomes and their biological plausibility, examined toxicological consistency across pathway layers, contributed to risk-assessment interpretation and regulatory translation, and reviewed and edited the manuscript. Y.P.G. conceived the overall research framework, supervised and coordinated the study, provided resources and funding support, guided the interpretation of the results, and critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (42322704, 42277222) and Guangdong Basic and Applied Basic Research Foundation (2023B1515020078). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The finalized model input data, endpoint labels, fixed cross-validation partitions and ablation inputs are available at https://doi.org/10.5281/zenodo.23475224 under accession 10.5281/zenodo.23475224, with the access and reuse conditions stated in the repository. The data underlying the supplementary tables are supplied as Supplementary Data 1, 2, 3a, 3b, 4a, 4b and 5–8. The reproduction package also includes archived starting data from redistributable sources, identified by third-party dataset versions or record accessions and file checksums. Source-specific reuse terms and the scope of each archived subset are stated in the package. The interactive AOPedex resource is available at https://aopedex.com.

Code availability

The core training and ablation scripts are available at https://doi.org/10.5281/zenodo.23475224 and archived at 10.5281/zenodo.23475224, under CC BY-SA 4.0. The package includes finalized data, fixed partitions, model configurations and instructions for training the supervised heads and rerunning the included ablations. It contains no data-acquisition or cleaning pipeline and no frontend or backend implementation of the interactive website.

Ethics

This computational study analyses existing datasets and published evidence. It does not involve new human participant research, identifiable personal data, live animal experiments or the collection of new biological samples. No ethics approval or informed consent was required.

AI assisted work

Literature retrieval and evidence-excerpt preparation were assisted by Kimi K3 Max, with evidence decisions made by human curators as described in Methods. Kimi K3 Max was also used for language polishing. The authors reviewed and edited all AI-assisted outputs and take full responsibility for the accuracy and integrity of the manuscript. No other AI tools were used.

References

1. Gore, A. C. et al. EDC-2: the Endocrine Society's second scientific statement on endocrine-disrupting chemicals. *Endocr. Rev.* **36**, E1–E150 (2015).

2. La Merrill, M. A. et al. Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification. *Nat. Rev. Endocrinol.* **16**, 45–57 (2020).

3. Kidd, K. A. et al. Collapse of a fish population after exposure to a synthetic estrogen. *Proc. Natl Acad. Sci. USA* **104**, 8897–8901 (2007).

4. Kidd, K. A. et al. Direct and indirect responses of a freshwater food web to a potent synthetic oestrogen. *Philos. Trans. R. Soc. B* **369**, 20130578 (2014).

5. Blanchfield, P. J. et al. Recovery of a wild fish population from whole-lake additions of a synthetic estrogen. *Environ. Sci. Technol.* **49**, 3136–3144 (2015).

6. McCallum, E. S., Kidd, K. A., Majewski, A. R. & Blanchfield, P. J. Whole-lake exposure to a widespread estrogenic pollutant leads to unanticipated reproductive behaviors in wild fish. *Curr. Biol.* **36**, 4513–4519.e3 (2026).

7. Desforges, J.-P. et al. Predicting global killer whale population collapse from PCB pollution. *Science* **361**, 1373–1376 (2018).

8. Lohmann, R. et al. Cross-cutting studies of per- and polyfluorinated alkyl substances (PFAS) in Arctic wildlife and humans. *Sci. Total Environ.* **954**, 176274 (2024).

9. Olker, J. H. et al. The ECOTOXicology Knowledgebase: a curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. *Environ. Toxicol. Chem.* **41**, 1520–1539 (2022).

10. McArdle, M. E. et al. Critical review of read-across potential in testing for endocrine-related effects in vertebrate ecological receptors. *Environ. Toxicol. Chem.* **39**, 739–753 (2020).

11. EFSA Panel on Plant Protection Products and their Residues (PPR) et al. Scientific opinion on the state of the science on pesticide risk assessment for amphibians and reptiles. *EFSA J.* **16**, e05125 (2018).

12. Bean, T. G. et al. Toxicological effects assessment for wildlife in the 21st century: review of current methods and recommendations for a path forward. *Integr. Environ. Assess. Manag.* **20**, 699–724 (2024).

13. Short, S. et al. Exploiting omic data to advance predictive ecotoxicology. *Environ. Sci. Technol.* **60**, 23642–23660 (2026).

14. Schoenfuss, H. L., Propper, C. R., Kolok, A. S. & Forbes, P. B. C. Terra (aqua) incognita: knowledge gaps in global ecotoxicology. *Environ. Toxicol. Chem.* **41**, 245–246 (2022).

15. Richard, A. M. et al. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem. Res. Toxicol.* **29**, 1225–1251 (2016).

16. Kramer, L. et al. Curated mode-of-action data and effect concentrations for chemicals relevant for the aquatic environment. *Sci. Data* **11**, 60 (2024).

17. Morrison, K. et al. Mapping meta-analyses on organochlorine pesticides reveals low methodological quality. *Nat. Sustain.* **8**, 1270–1279 (2025).

18. Wang, B. et al. Building the foundations of AI-driven toxicology: how to use fragmented data for mechanism-based human health risk assessment. *Environ. Sci. Technol.* **60**, 26908–26921 (2026).

19. Dwyer, F. J. et al. Assessing contaminant sensitivity of endangered and threatened aquatic species: part I. Acute toxicity of five chemicals. *Arch. Environ. Contam. Toxicol.* **48**, 143–154 (2005).

20. Liang, W., Li, J., Wang, X., Giesy, J. P. & Zhao, X. Machine learning-driven cross-species toxicity prediction for advancing ecologically relevant PFAS water quality criteria. *Environ. Sci. Technol.* **59**, 25688–25702 (2025).

21. Wang, R., Yi, S., Shan, G. & Zhu, L. Integration of transfer learning and multitask learning to predict the potential of per/polyfluoroalkyl substances in activating multiple nuclear receptors associated with hepatic lipotoxicity. *Environ. Sci. Technol.* **59**, 23727–23738 (2025).

22. Naser, M. Z. A review of machine learning with small and limited data. *J. Big Data* **13**, 18 (2026).

23. Schlender, T. et al. The bigger fish: a comparison of meta-learning QSAR models on low-resourced aquatic toxicity regression tasks. *Environ. Sci. Technol.* **57**, 17818–17830 (2023).

24. Zhao, C., Wang, J., Bo, X. & He, S. ToxiSpecies: task-aware meta-learning for cross-species modeling of acute chemical toxicity under distribution shift. *J. Chem. Inf. Model.* **66**, 9226–9235 (2026).

25. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).

26. Ankley, G. T. & Gray, L. E. Cross-species conservation of endocrine pathways: a critical analysis of Tier 1 fish and rat screening assays with 12 model chemicals. *Environ. Toxicol. Chem.* **32**, 1084–1087 (2013).

27. LaLone, C. A. et al. Sequence alignment to predict across species susceptibility (SeqAPASS): a web-based tool for addressing the challenges of cross-species extrapolation of chemical toxicity. *Toxicol. Sci.* **153**, 228–245 (2016).

28. von Rueden, L. et al. Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.* **35**, 614–633 (2023).

29. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).

30. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* **55**, 66 (2023).

31. Knapen, D. et al. Adverse outcome pathway networks I: development and applications. *Environ. Toxicol. Chem.* **37**, 1723–1733 (2018).

32. OECD. *Guidance Document for the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and Assessment (IATA)* (OECD Publishing, Paris, 2017).

33. Hartung, T. & Rovida, C. Mechanistic read-across comes of age: a comparative appraisal of EFSA 2025 guidance, ECHA's RAAF, and good read-across practice. *Front. Toxicol.* **7**, 1690491 (2025).

34. Haigis, A.-C. et al. Cross-species applicability of an adverse outcome pathway network for thyroid hormone system disruption. *Toxicol. Sci.* **195**, 1–27 (2023).

35. Xin, L. et al. Knowledge-based machine learning for predicting and understanding the androgen receptor (AR)-mediated reproductive toxicity in zebrafish. *Environ. Int.* **191**, 108995 (2024).

36. Tan, H. et al. Causality-integrated graph learning for multi-endpoint toxicity prediction. *Proc. Natl Acad. Sci. USA* **123**, e2608919123 (2026).

37. Wang, Y., Song, L., Wu, Y., Fantke, P. & Li, S. Using multimodal large language models for extracting structured life-cycle inventory data from scientific literature. *Environ. Sci. Technol.* **60**, 27792–27802 (2026).

38. Ankley, G. T. et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* **29**, 730–741 (2010).

39. Ives, C., Campia, I., Wang, R.-L., Wittwehr, C. & Edwards, S. Creating a structured adverse outcome pathway knowledgebase via ontology-based annotations. *Appl. In Vitro Toxicol.* **3**, 298–311 (2017).

40. Dix, D. J. et al. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **95**, 5–12 (2007).

41. McRobb, F. M., Sahagún, V., Kufareva, I. & Abagyan, R. In silico analysis of the conservation of human toxicity and endocrine disruption targets in aquatic species. *Environ. Sci. Technol.* **48**, 1964–1972 (2014).

42. OECD. *Revised Guidance Document 150 on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption* (OECD Publishing, Paris, 2018).

43. Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 623–631 (ACM, 2013).

44. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: a unified framework for machine learning interpretability. Preprint at https://doi.org/10.48550/arXiv.1909.09223 (2019).

45. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326 (2025).

46. von Borries, K. et al. Uncertainty-aware machine learning to predict non-cancer human toxicity for the global chemicals market. *Nat. Commun.* **17**, 647 (2026).

47. Zhang, K., Pan, R., Shi, H. & Zhao, Y. Beyond oxidative stress: rethinking mechanistic inference in big data-driven adverse outcome pathway construction. *Environ. Sci. Technol.* **60**, 24985–24988 (2026).

48. OECD. *Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways* (OECD Publishing, Paris, 2018).

49. Tong, W. et al. Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **112**, 1249–1254 (2004).

50. Blomberg, S. P., Garland, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).

51. Shang, Y. et al. Allometric equations for aboveground biomass estimation in natural forest trees: generalized or species-specific? *Diversity* **17**, 493 (2025).

52. Brown, B. P. A generalizable deep learning framework for structure-based protein–ligand affinity ranking. *Proc. Natl Acad. Sci. USA* **122**, e2508998122 (2025).

53. Krasowski, M. D., Yasuda, K., Hagey, L. R. & Schuetz, E. G. Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). *Nucl. Recept.* **3**, 2 (2005).

54. Cuvillier-Hot, V. & Lenoir, A. Invertebrates facing environmental contamination by endocrine disruptors: novel evidences and recent insights. *Mol. Cell. Endocrinol.* **504**, 110712 (2020).

55. Matthiessen, P., Wheeler, J. R. & Weltje, L. A review of the evidence for endocrine disrupting effects of current-use chemicals on wildlife populations. *Crit. Rev. Toxicol.* **48**, 195–216 (2018).

56. IUCN. *The IUCN Red List of Threatened Species*. https://www.iucnredlist.org.

57. Mizoguchi, B., Topping, N. E., Lavin, A. M. & Valenzuela, N. Cadmium ecotoxic effects on embryonic Dmrt1 and aromatase expression in Chrysemys picta turtles may implicate changes in DNA methylation. *Genes* **13**, 1318 (2022).

58. Shen, J. et al. Loss- and gain-of-function analyses reveal the essential role of Cyp19a1 in ovarian determination of the red-eared slider turtle. *Genetics* **224**, iyad041 (2023).

59. Zhao, Y., Lin, M.-C. A., Mock, A., Yang, M. & Wayne, N. L. Kisspeptins modulate the biology of multiple populations of gonadotropin-releasing hormone neurons during embryogenesis and adulthood in zebrafish (Danio rerio). *PLoS ONE* **9**, e104330 (2014).

60. Pasquier, J. et al. Looking for the bird Kiss: evolutionary scenario in sauropsids. *BMC Evol. Biol.* **14**, 30 (2014).

61. Robledo, D. A. R. et al. Ecological and structural drivers underlying species-specific estrogen receptor α responses to estrogenic chemicals in mammals. *Environ. Sci. Technol.* **60**, 25279–25292 (2026).

62. Greaves, A. K., Letcher, R. J., Sonne, C. & Dietz, R. Brain region distribution and patterns of bioaccumulative perfluoroalkyl carboxylates and sulfonates in East Greenland polar bears (Ursus maritimus). *Environ. Toxicol. Chem.* **32**, 713–722 (2013).

63. Tang, L. et al. Chemicals of environmental concern as inhibitors of human placental 3β-hydroxysteroid dehydrogenase 1 and aromatase: screening and docking analysis. *Chem. Biol. Interact.* **368**, 110243 (2022).

64. Xu, Y. et al. Chain length-dependent inhibition of human and rat placental aromatase by benzalkonium disinfectants: experimental, quantitative structure–activity relationship, and in silico docking insights. *Environ. Sci. Technol.* **59**, 20990–21002 (2025).

65. Kang, J. S., Ahn, T.-G. & Park, J.-W. Perfluorooctanoic acid (PFOA) and perfluooctane sulfonate (PFOS) induce different modes of action in reproduction to Japanese medaka (Oryzias latipes). *J. Hazard. Mater.* **368**, 97–103 (2019).

66. Veldhoen, N. et al. The bactericidal agent triclosan modulates thyroid hormone-associated gene expression and disrupts postembryonic anuran development. *Aquat. Toxicol.* **80**, 217–227 (2006).

67. Li, H. et al. Triclocarban and triclosan inhibit human aromatase via different mechanisms. *Biomed Res. Int.* **2017**, 8284097 (2017).

68. Chen, J., Meng, T., Li, Y., Gao, K. & Qin, Z. Effects of triclosan on gonadal differentiation and development in the frog Pelophylax nigromaculatus. *J. Environ. Sci.* **64**, 157–165 (2018).

69. Jain, S. & Wallace, B. C. Attention is not explanation. In *Proceedings of NAACL-HLT 2019, Vol. 1* 3543–3556 (Association for Computational Linguistics, 2019).

70. Sundararajan, M. & Najmi, A. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning* **119**, 9269–9278 (PMLR, 2020).

71. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should I trust you?": explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).

72. Heskes, T., Sijben, E., Bucur, I. G. & Claassen, T. Causal Shapley values: exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems* **33**, 4778–4789 (2020).

73. Yang, L., Zeng, J., Gao, N., Zhu, L. & Feng, J. Predicting the metal mixture toxicity with a toxicokinetic–toxicodynamic model considering the time-dependent adverse outcome pathways. *Environ. Sci. Technol.* **58**, 3714–3725 (2024).

74. Abbate, E. et al. Screening chemical alternatives with limited ecotoxicity data to support safe and sustainable by design innovation. *Environ. Sci. Technol.* **60**, 27104–27114 (2026).

75. Yan, Z. et al. Bridging exposure and mechanism: a mechanism-weighted ecological risk assessment reveals divergent priorities for endocrine-disrupting chemicals in China. *Environ. Sci. Technol.* https://doi.org/10.1021/acs.est.6c10115 (2026).

76. Bart, S. et al. How to analyse and account for interactions in mixture toxicity with toxicokinetic-toxicodynamic models. *Sci. Total Environ.* **843**, 157048 (2022).

77. LeCun, Y. A path towards autonomous machine intelligence. Preprint at https://openreview.net/forum?id=BZ5a1r-kVsf (2022).

78. Ding, J. et al. Understanding world or predicting future? A comprehensive survey of world models. *ACM Comput. Surv.* **58**, 57 (2026).

79. Klindt, D., LeCun, Y. & Balestriero, R. When does LeJEPA learn a world model? Preprint at https://doi.org/10.48550/arXiv.2605.26379 (2026).

80. Mishra, G. Ecology is not yet ready for AI—and why that matters. *Proc. Natl Acad. Sci. USA* **123**, e2533976123 (2026).

81. Zhu, J.-J., Yang, M. & Ren, Z. J. Machine learning in environmental research: common pitfalls and best practices. *Environ. Sci. Technol.* **57**, 17671–17689 (2023).

82. Kapoor, S. et al. REFORMS: consensus-based recommendations for machine-learning-based science. *Sci. Adv.* **10**, eadk3452 (2024).

83. Wang, Y.-Q. et al. Auditable automation of activated sludge modeling for wastewater treatment diagnosis using LLM-agents. *Environ. Sci. Technol.* **60**, 27879–27892 (2026).

84. Chung, E., Russo, D. P., Aleksunes, L. M., Warner, G. R. & Zhu, H. Hierarchical mechanistic modeling of complex toxicity endpoints from public concentration–response data. *Environ. Sci. Technol.* **60**, 2539–2555 (2026).

85. Elliott, J. H. et al. Living systematic review: 1. Introduction—the why, what, when, and how. *J. Clin. Epidemiol.* **91**, 23–30 (2017).

86. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).

87. Gottweis, J. et al. Accelerating scientific discovery with Co-Scientist. *Nature* (2026).

88. Ghareeb, A. E. et al. A multi-agent system for automating scientific discovery. *Nature* (2026).

89. Minot, M. et al. Advancing chemical reaction prediction in data-scarce drug discovery with active and geometric deep learning. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-026-01063-0 (2026).

90. Lai, H. et al. Interformer: an interaction-aware model for protein-ligand docking and affinity prediction. *Nat. Commun.* **15**, 10223 (2024).

91. VanderWeele, T. J. & Ding, P. Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.* **167**, 268–274 (2017).

92. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).

93. Doering, J. A. et al. Differential sensitivity to in vitro inhibition of cytochrome P450 aromatase (CYP19) activity among 18 freshwater fishes. Toxicol. Sci. 170, 394–403 (2019). https://doi.org/10.1093/toxsci/kfz115

Supplementary Files

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

- 02SupplementaryInformation.docx
- SupplementaryData.zip
- SourceData.zip