Multimodal Cross-Attentive Graph-Based Framework for Predicting *In Vivo* Endocrine Disruptors

Eder Soares de Almeida Santos,[1] Gustavo Felizardo Santos Sandes,[1] Artur Christian Garcia da Silva,[2] Holli-Joi Martin,[3] Eugene N. Muratov,[3] Rodolpho de Campos Braga,[4] Bruno Junior Neves[1,*]

[1] Laboratory of Cheminformatics, Faculty of Pharmacy, Federal University of Goiás, Goiânia, 74605-170, Brazil

[2] Laboratory of Education and Research in *In Vitro* Toxicology, Faculty of Pharmacy, Universidade Federal de Goiás, Goiás, 74605-170, Brazil

[3] Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, North Carolina, 27599-7360, USA

[4] InsilicAll Ltda., São Paulo, 04571-010, Brazil

*Author for correspondence: brunoneves@ufg.br

ABSTRACT

Endocrine hazard assessment needs models that are accurate and mechanistically transparent. We present a multimodal cross-attentive graph framework that fuses molecular graphs with adverse-outcome-pathway (AOP)–anchored assay signals to predict organism-level outcomes in the OECD Hershberger and uterotrophic assays. In Tier-1, multitask GNNs learn ER/AR molecular-initiating and key events across 46 ToxCast/Tox21 assays. In Tier-2, a cross-attentive multimodal GNN integrates Tier-1 pathway signals with molecular graphs, achieving AUROC = 0.90 (Hershberger) and 0.96 (uterotrophic). External validation on literature compounds showed 84% concordance (Hershberger 15/18; uterotrophic 22/26). Bidirectional cross-attention links molecular substructures to pathway assays and vice-versa, while counterfactual perturbations rank assays and structural motifs most responsible for each decision. The framework couples high accuracy with assay-traceable explanations, supporting targeted testing within the Integrated Approaches.

KEYWORDS

Endocrine disruption; deep learning; adverse outcome pathway; estrogen; androgen.

1. INTRODUCTION

Endocrine-disrupting chemicals (EDCs) comprise a group of exogenous substances that interfere with hormone signaling pathways, particularly through estrogen (ER) and androgen (AR) receptors mechanisms,[1,2] leading to well-documented adverse outcomes ranging from reproductive disorders to hormone-dependent cancers.[3–7] International evaluations and economic analyses estimate that such exposures contribute to widespread disease at the population level and generate annual costs amounting to hundreds of billions of euros in the European Union.[8]

Assessment of chemical hazards has historically relied on *in vivo* assays established under the Organization for Economic Co-operation and Development (OECD) test guidelines to capture AR- and ER-mediated mechanisms. The Hershberger assay (OECD test no. 441), conducted in castrated male rats supplemented with testosterone propionate, provides direct evidence of AR agonism or antagonism through changes in male accessory sex organ weights. Similarly, the uterotrophic assay (OECD test no. 440), performed in immature or ovariectomized female rodents, monitors ER modulation by measuring uterine weight and histological changes. While these assays serve as regulatory benchmarks for identifying endocrine activity, they have limited throughput, high costs, and ethical concerns associated with the use of animals.[9]

In addition, high-throughput screening strategies have been established for assessing EDCs. The U.S. Environmental Protection Agency's (EPA) ToxCast program and the interagency Tox21 collaboration have generated extensive *in vitro* bioactivity data across ~10,000 chemicals,[10–12] including dedicated assay batteries for ER and AR pathways.[13,14] These resources provide unprecedented mechanistic insight for regulatory prioritization but remain constrained by limited metabolic competence and the challenge of translating pathway perturbations into complex *in vivo* outcomes.[15,16] Building on these efforts, the U.S. EPA implemented the Endocrine Disruptor Screening Program (EDSP) to determine potential endocrine effects in humans and wildlife. The program was structured in two tiers: Tier-1 was designed to identify potential EDCs through a battery of *in vitro* and *in vivo* assays (including the OECD Hershberger and uterotrophic tests); and Tier-2, intended to characterize dose-response relationships and adverse outcomes in more comprehensive animal studies.[17,18] Despite its ambitious scope, the EDSP has progressed slowly, with Tier-1 testing completed for only a limited number of chemicals.

Considering these limitations, computational toxicology has emerged as a promising approach to enhance the throughput of chemical assessments.[19] Within computational toxicology, machine learning approaches leverage large-scale biological, chemical, and mechanistic data to uncover non-

linear patterns underlying endocrine disruption,[20–27] thereby enabling regulatory prioritization while reducing reliance on animal testing. These approaches have been increasingly adopted into Integrated Approaches to Testing and Assessment (IATA) frameworks as complementary tools to conventional assays, offering a scalable strategy to screen the thousands of chemicals that remain untested for endocrine activity.[28,29]

In recent years, graph neural networks (GNNs)[30,31] have expanded the landscape of predictive toxicology. These models rely on molecular graphs, where nodes represent atoms, edges represent bonds, and attributes (such as atom type, chirality, and bond order) are encoded as features. GNNs offer adaptable architectures for modeling complex molecular interactions and endocrine-related responses. However, current applications in computational toxicology rely primarily on molecular structure information, with underutilization of biological pathway data. Despite promising advances, these studies remain limited in number and scope, focusing mainly on *in vitro* AR- and ER-mediated effects.[32–34] Extending these efforts toward more comprehensive predictions, particularly those aligned with organism-level outcomes, requires integrative modeling strategies that combine molecular graph representations with adverse outcome pathways and assay-derived information. Such approaches enable the capture of receptor-level interactions while also accounting for downstream biological responses.

In this study, we introduce a multimodal cross-attentive graph framework that uniquely integrates chemical structure information with adverse outcome pathway (AOP)-based predictions to model *in vivo* endocrine disruption. The approach was developed in two tiers: first, multitask GNNs were trained on 46 *in vitro* assays from the ToxCast/Tox21 program to capture AR- and ER-mediated molecular initiating events (MIEs) and key events (KEs); second, the resulting predictions and logits—rather than the raw *in vitro* labels—were embedded within a pathway graph and integrated with molecular graphs through cross-attentive multimodal GNNs to predict outcomes of the OECD Hershberger (AUROC = 0.90) and uterotrophic (AUROC = 0.96) assays.

Beyond predictive performance, the multimodal framework provides explainability through bidirectional cross-attention, which links molecular substructures to pathway assays and, conversely, pathway nodes to their most influential molecular features. This dual mapping elucidates how structural motifs propagate into AR- and ER-mediated perturbations. As an additional interpretability strategy, counterfactual perturbations were employed to generate alternative scenarios, enabling the evaluation of how molecular modifications or pathway disruptions may alter downstream outcomes.

2. RESULTS AND DISCUSSION

2.1. Tier-1

Tier-1 modeling was designed to predict *in vitro* responses across AR- and ER-mediated assays, establishing a mechanistic basis for downstream *in vivo* predictions. Initially, a dataset of 7,957 chemicals tested in 46 *in vitro* assays from the U.S. EPA ToxCast/Tox21 program (invitroDB v4.1) was compiled within AR- and ER-mediated AOPs (Figure 1). These assays capture a structured sequence of biological events, linking MIEs such as receptor binding to successive KEs (e.g., transcriptional regulation, gene expression, and cell proliferation) that ultimately culminate in an adverse outcome (AO) at the organismal level.

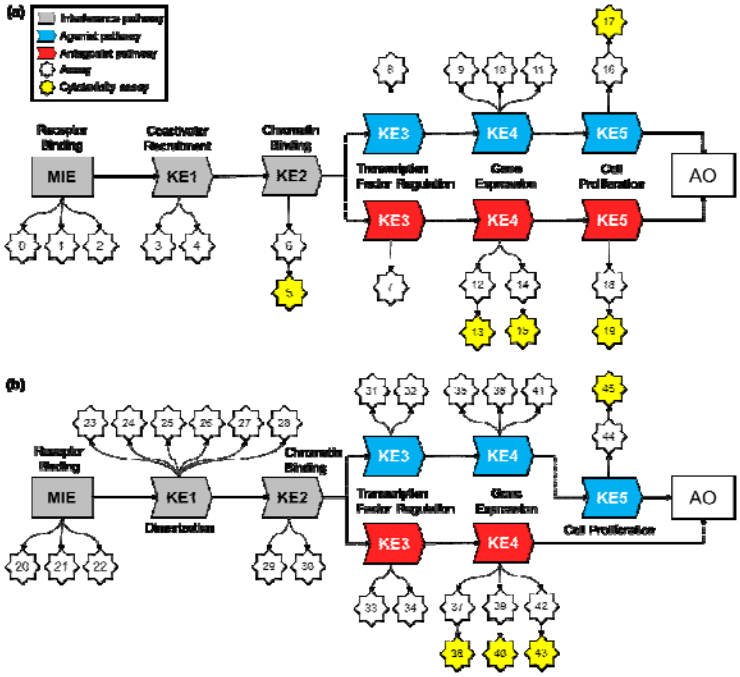


Figure 1. Graphical representation of the (**a**) AR- and (**b**) ER-mediated AOPs based on Tox21/ToxCast assays.

As shown in Figure 1a, the AR-mediated AOP included 20 *in vitro* assays, comprising three receptor binding assays (MIE), two co-regulator recruitment assays (KE1), one chromatin binding assay (KE2), two transcription factor regulation assays (KE3), five gene expression assays (KE4), two cell proliferation assays (KE5), and five cytotoxicity assays. The ER-mediated AOP (Figure 1b) encompassed 26 *in vitro* assays, including three receptor binding assays (MIE), six dimerization and co-regulator recruitment assays (KE1), two chromatin binding assays (KE2), four transcription factor activity assays (KE3), six assays for gene expression (KE4), one cell proliferation assay (KE5), and four cytotoxicity assays. Cytotoxicity assays were incorporated into both AOPs to

provide critical context for interpreting agonistic and antagonistic responses, as they help discriminate true receptor-mediated activity from nonspecific effects driven by generalized cell stress.[35,36]

These assays were subsequently employed to train multitask learning models aimed at predicting AR- and ER-mediated responses. Modeling this dataset poses a substantial challenge, as the average imbalance ratio across training, validation, and test splits was 0.86, reflecting the dominance of inactives in the dataset (Figure 2a). To address this, four GNN architectures were explored (MPNN, GAT, GIN, and AttentiveFP), each optimized with virtual node embeddings, skip connections, and Jumping Knowledge mechanisms to improve chemical semantic representation and mitigate oversmoothing and oversquashing during training (Figure 2b). Performance evaluation revealed global AUROC values of 0.83 for MPNN, 0.81 for GAT, 0.78 for GIN, and 0.79 for AttentiveFP, with the multitask MPNN emerging as the most effective backbone for Tier-1 predictions (Figure 2c).

To further address heterogeneity across tasks, we implemented a homoscedastic uncertainty weighting framework to adaptively balance loss contributions during multitask learning. In this approach, each task is assigned a trainable variance parameter that downweights highly uncertain tasks, preventing them from dominating the optimization process. Meanwhile, a regularization term penalizes inflated variance estimates to avoid trivial solutions. As shown in Figure 2d, incorporating uncertainty weighting consistently improved performance across all architectures when compared with their default counterparts. The MPNN, which already achieved the best baseline performance (AUROC = 0.83), increased to 0.84 (+1.2%). The GAT exhibited the most significant relative gain, from 0.81 to 0.84 (+3.7%), while both GIN and AttentiveFP improved from 0.79 to 0.82 (+3.8%). These results indicate that homoscedastic uncertainty weighting provides a systematic advantage across diverse GNN backbones, with the strongest effect observed in models with lower baseline performance.

Figure 2e shows that the uncertainty-weighted MPNN maintained high accuracy across splits (ACC ≈0.89–0.91). However, data imbalance resulted in high specificity (SP = 0.98) but low recall (0.27–0.42), primarily reflecting the dominance of inactives. Consistently, G-mean values remained low (0.52–0.64) across splits, as highlighted in the accompanying box plots. The distribution of predicted probabilities (Figure 2f) further illustrates this bias, with inactives clustering near zero and most actives falling below the standard 0.5 decision boundary, highlighting that this default threshold is not optimal in such an imbalanced setting.

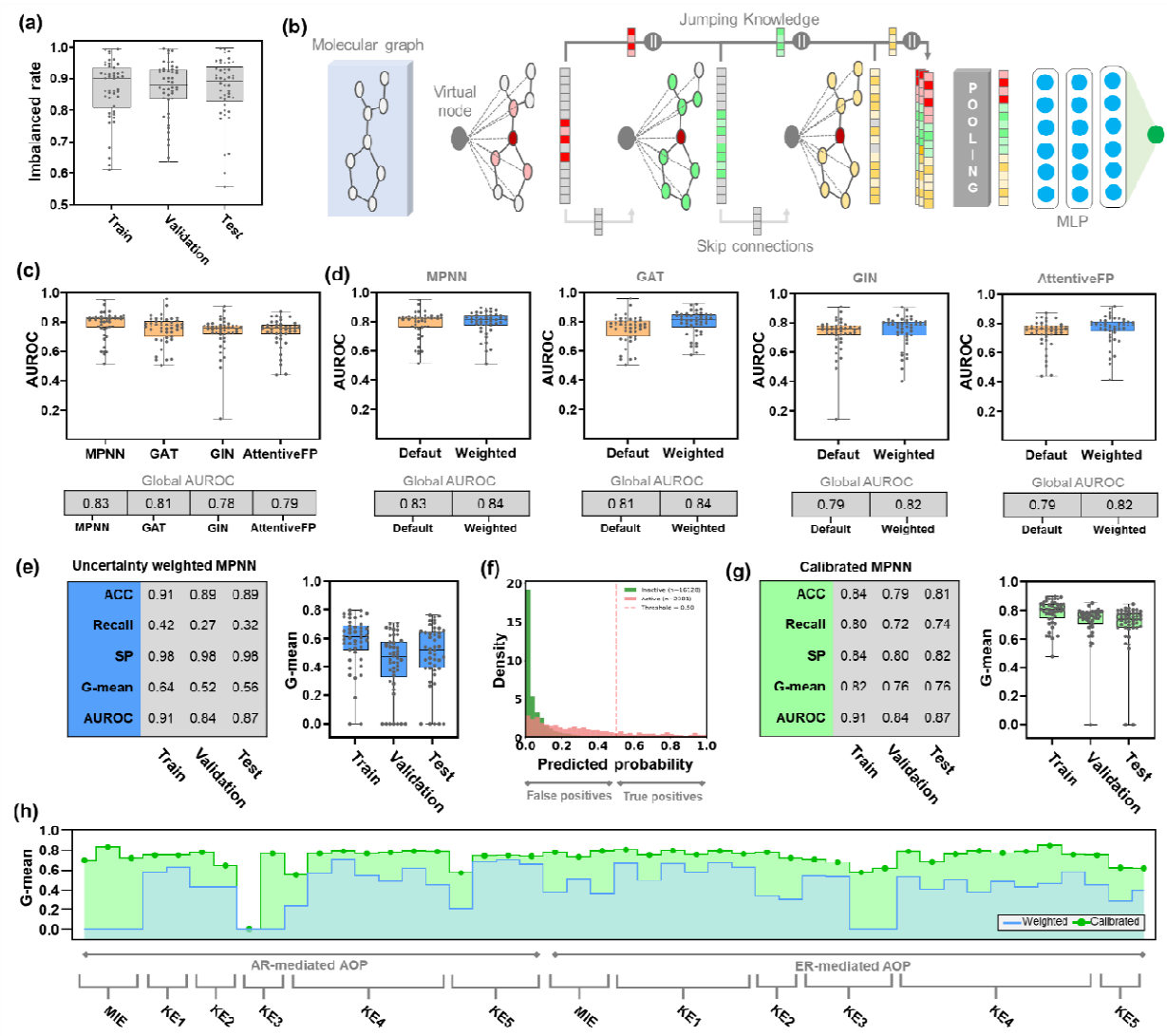


Figure 2. Statistical characteristics of multitask GNNs for predicting AR- and ER-mediated assay outcomes. (**a**) Imbalance ratios across training, validation, and test splits. (**b**) Schematic representation of multitask GNN architectures with virtual node, skip connections, and Jumping Knowledge mechanisms. (**c**) Global AUROC values of four GNN backbones. (**d**) Comparison of default vs. uncertainty-weighted loss across architectures. (**e**) Global performance metrics of the uncertainty-weighted MPNN and box plots of G-mean across tasks. (**f**) Predicted probability distributions showing class imbalance and suboptimality of the 0.5 threshold. (**g**) Global performance metrics of the calibrated MPNN with task-specific thresholds and box plots of G-mean across tasks. (**h**) Task-level G-mean before and after calibration across AR- and ER-mediated AOPs.

Probabilistic calibration[37] was then applied to the best-performing model (MPNN) using task-specific thresholds optimized by G-mean. This calibration effectively shifts the decision boundaries on a task-by-task basis, moving them away from the default 0.5 threshold toward values that balance

recall and SP. As a result, most active-class probability distributions undergo a leftward shift, recovering previously misclassified actives while preserving overall discrimination. As shown in Figure 2g, this strategy markedly increased recall (0.72–0.80) while maintaining balanced specificity (≈0.80–0.84). Consequently, G-mean values improved to 0.76–0.82 across splits, corresponding to relative gains of +28% (train), +46% (validation), and +36% (test), as illustrated in the box plots. At the individual assay level (Figure 2h), task-wise calibration yielded consistent improvements in G-mean across most endpoints within both AR- and ER-mediated AOPs. Only a few tasks showed limited or no gain, which is consistent with the high imbalance ratios and sparse representation of minority-class samples across the data splits in these assays.

2.2. Tier-2

Tier-2 modeling was designed to predict *in vivo* endocrine outcomes, specifically the Hershberger and uterotrophic assays, by integrating chemical features with pathway-derived predictions from Tier-1. The Hershberger dataset[38] comprised 136 compounds, with activity defined by androgen-dependent changes in the weights of accessory sex tissues. The uterotrophic dataset[39] comprised 118 compounds, where estrogenic activity was determined by measuring uterine weight changes in immature or ovariectomized female rodents. To expand the limited sample size of both assays, data augmentation was applied by incorporating experimentally confirmed inactives, consistently identified across all ToxCast/Tox21 assays (Hershberger: 123 compounds, uterotrophic: 136 compounds), as well as 13 non-toxicants from EDSP Tier-1 assessments. As shown in Figure 3a, the Hershberger dataset consisted of 69 positive and 203 negative samples, while the uterotrophic dataset comprised 64 positive and 203 negative samples. Notably, both datasets provided broad coverage of the chemical space represented in the *in vitro* ToxCast/Tox21 assays (Figure 3b), indicating that the compounds used for Tier-2 predictions are well distributed across the diversity of tested chemicals.

To capture mechanistic context, pathway graphs were generated from AR- and ER-mediated AOPs, with nodes representing individual Tier-1 assays and edges reflecting their hierarchical position and co-occurrence patterns. These pathway graphs were integrated with molecular graphs through a multimodal GNN architecture (Figure 3c), where message-passing layers (AttentiveFP, GAT, GIN, and MPNN) processed each graph separately and bidirectional cross-attention blocks enabled information exchange between them. The refined molecular and pathway embeddings were

then pooled, concatenated, and passed to a final classifier, providing joint representations optimized for predicting *in vivo* outcomes.

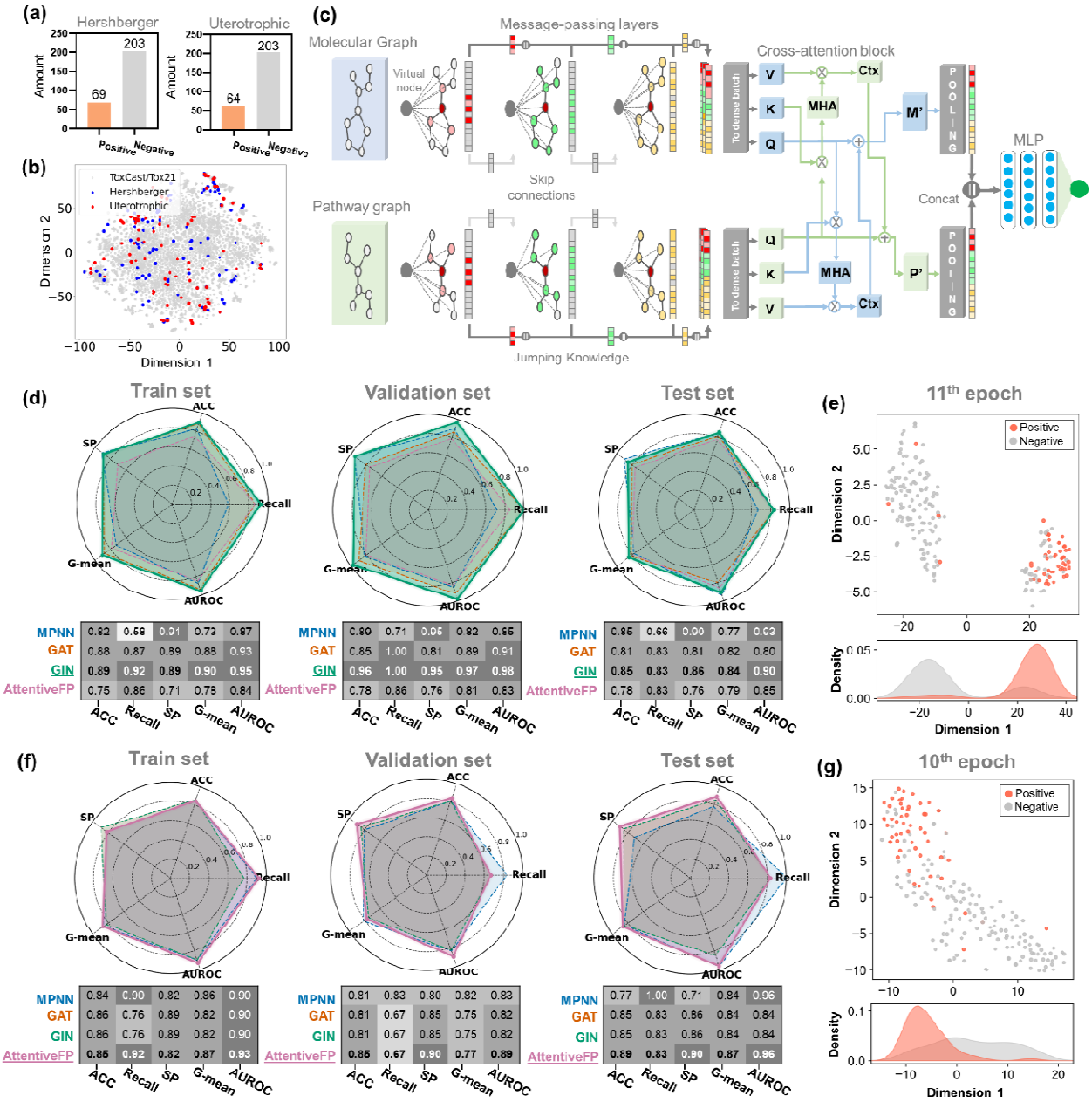


Figure 3. Statistical characteristics of multimodal GNN framework for predicting outcomes in Hershberger and uterotrophic assays. (**a**) Distribution of positive and negative samples in Hershberger and uterotrophic datasets after augmentation with consistently inactive compounds. (**b**) t-SNE projection of compounds showing broad coverage of the ToxCast/Tox21 chemical space. (**c**) Schematic representation of multimodal GNN architecture integrating molecular and pathway graphs through message-passing, cross-attention, and pooling. Predictive performance of

multimodal GNN backbones (MPNN, GAT, GIN, AttentiveFP) on the (**d**) Hershberger and (**f**) uterotrophic assays across training, validation, and test sets. t-SNE embeddings of (**e**) Hershberger (GIN-based model, 11th epoch) and (**g**) uterotrophic predictions (AttentiveFP-based model, 10th epoch) showing clear separation of positive and negative compounds with corresponding KDE distributions.

All multimodal GNN architectures demonstrated strong predictive performance for the Hershberger assay, with AUROC values consistently above 0.80 across validation and test (Figure 3d). Among them, the GIN-based model (probability threshold = 0.40) provided the best overall results, achieving AUROCs of 0.98 and 0.90, respectively. This model also reached the highest accuracy (0.96–0.85) and the best balance between Recall and SP, as reflected by G-mean values of 0.97–0.84. This superior performance was further supported by the t-SNE projection of learned embeddings (Figure 3e), which revealed a clear separation between positive (orange) and negative (gray) compounds. The corresponding KDE distributions confirmed minimal overlap between the two classes, highlighting the model's ability to generate well-structured latent representations that discriminate against AR-mediated responses. Multimodal GNNs also showed consistently high predictive performance in the uterotrophic assay, achieving AUROC values greater than 0.82 across training, validation, and test sets (Figure 3f). The AttentiveFP-based model (probability threshold = 0.50) yielded the strongest results, reaching AUROCs of 0.89 and 0.96 for the validation and test sets, respectively. This model also maintained high accuracy (0.85–0.89) and delivered the most balanced trade-off between recall and specificity, as reflected by G-mean values of 0.77–0.87. The t-SNE projections of learned embeddings (Figure 3g) revealed distinct clustering of positive (orange) and negative (gray) compounds, with minimal overlap in KDE density, reinforcing the strength of AttentiveFP in modeling ER-mediated responses.

2.3. Ablation study

Ablation analysis was next performed to evaluate the contribution of each type of information (i.e., Tier-1 predictions, molecular graphs, and pathway graphs) to the predictive performance of Tier-2 models (Figure 4). Initially, the top-K algorithm was developed to assess whether Tier-1 predictions alone were sufficient for *in vivo* extrapolation. This approach classifies a compound as positive when the number of activated levels (MIEs and KEs) within the AOP exceeds a predefined threshold, reflecting the principle that progression across multiple key events increases the likelihood of an adverse outcome.[40,41] As shown in Figure 4a, this approach yielded low

performance, with an AUROC of 0.61 ± 0.14 for the Hershberger assay (−0.27 AUROC) and 0.63 ± 0.20 for the uterotrophic assay (−0.27 AUROC). These results suggest that *in vitro* outputs, when isolated, lack the resolution necessary to capture organism-level responses.

Algorithm: Top-K algorithm for *in vivo* endocrine disruption classification

Input: Tier-1 predictions P, thresholds τ
Initialize selected tasks T
while session active **do**
  Binarize $B = [P \geq \tau]$
  Apply viability filter $\rightarrow B'$
  Identify effective tasks E
  Count active levels $L = \left|\{\ell \in \text{levels} \mid \exists_j \in E : B'_{:,j} = 1\}\right|$
  if $L \geq L_{min}$ *or* $\sum_{j \in E} B'_{:,j} \geq k$ **then**
    Assign toxic
  else non-toxic

Subsequently, the contributions of molecular and pathway graphs were examined independently by deactivating each arm of the multimodal GNN framework. Models restricted to molecular graphs (Figure 4b) showed limited performance, with mean AUROC values of 0.68 ± 0.16 for the Hershberger assay (−0.20 AUROC) and 0.60 ± 0.21 for the uterotrophic assay (−0.30 AUROC), compared to the multimodal design (0.88 ± 0.06 and 0.90 ± 0.06, respectively). Pathway graph models (Figure 4c) yielded relevant prediction performances, achieving 0.82 ± 0.06 for the Hershberger assay (−0.06 AUROC) and 0.80 ± 0.10 for the uterotrophic assay (−0.10 AUROC); however, they underperformed relative to the multimodal framework. Overall, the combination of molecular and pathway graphs in the multimodal framework delivered the most accurate Tier-2 predictions.

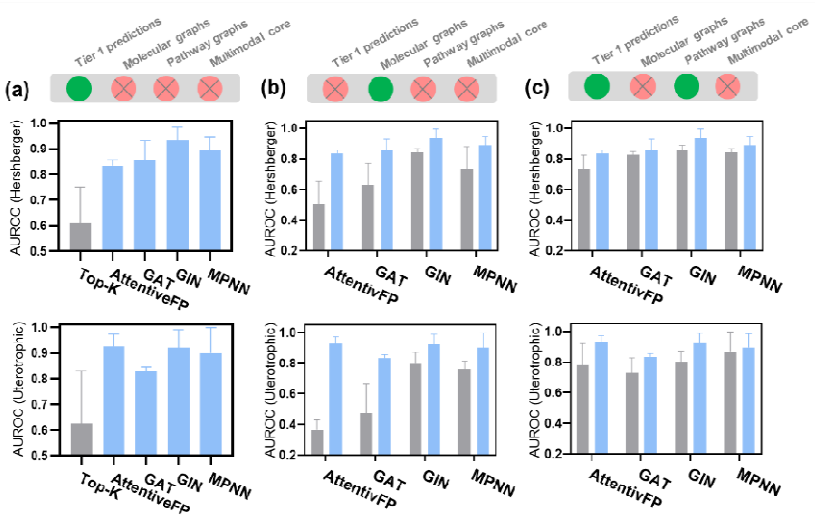


Figure 4. Ablation analysis of multimodal GNNs for predicting outcomes in Hershberger and uterotrophic assays. (**a**) AUROC evaluation of the top-K rule-based algorithm using Tier-1 predictions to classify compounds according to the number of activated AOP levels. (**b**) AUROC evaluation of the framework restricted to molecular graphs, isolating the contribution of chemical structure information. (**c**) AUROC evaluation of the framework restricted to pathway graphs, isolating the contribution of AOP-derived features. Error bars represent standard deviations across data splits.

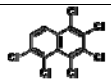
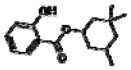
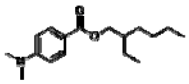
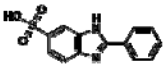
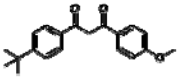
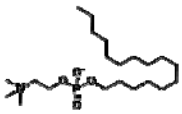
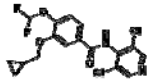
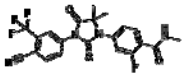
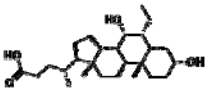
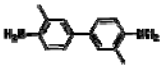
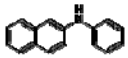
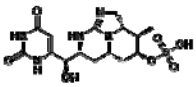
2.4. External validation with literature compounds

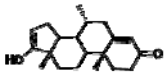
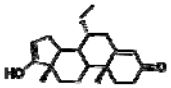
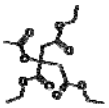
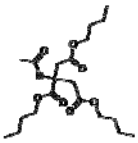
To further interrogate the robustness and translational value of the Tier-2 models, an external validation was performed using compounds identified through a systematic literature review. Equivocal or conflicting reports, as well as mixtures and compounds associated with systemic effects such as changes in body weight, were excluded from the analysis. This evaluation included 19 chemicals with androgen-related outcomes directly relevant to the Hershberger assay and 26 chemicals with estrogen-related outcomes relevant to the uterotrophic assay (Tables 1 and 2). The selected compounds encompassed both substances tested in OECD guideline studies and others with well-documented *in vivo* outcomes strongly indicative of the same endocrine endpoints, thereby providing a stringent and diverse benchmark for model assessment.

Application of the Tier-2 framework to this independent dataset revealed a high degree of concordance between predicted and observed outcomes. The Hershberger model correctly classified 15 of 18 compounds (ACC = 83%), while the uterotrophic model accurately predicted 22 of 26 cases (ACC = 84%). Among the compounds evaluated, specific examples illustrate the predictive

capacity of the Tier-2 framework. Padimate-O (CAS no. 21245-02-3), a UV filter, was correctly identified as positive by the model (probability = 0.70). Consistent with experimental evidence, *in vivo* studies report decreases in glans penis, bulbocavernous muscle complex, and seminal vesicle weights in rats exposed orally at 1000 mg/kg/day for 10 days.[42] In contrast, ensulizole (CAS no. 27503-81-7), another UV filter, was correctly predicted as negative (probability = 0.001), in agreement with Hershberger assays, which showed no alterations in androgen-dependent tissues up to the limit dose of 1000 mg/kg/day.[43]

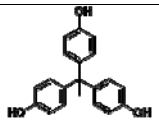
Table 1. External validation of the Hershberger model with chemicals identified from the literature.

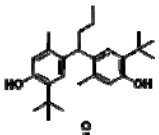
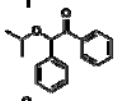
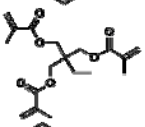
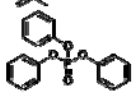
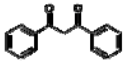
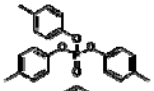
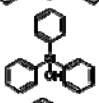
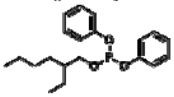
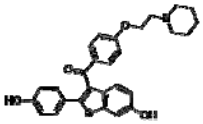
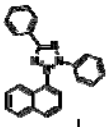
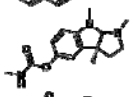
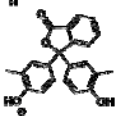
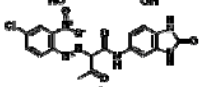
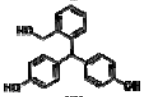
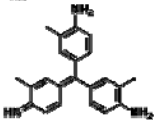
Compound (CAS)	**usage**	**Structure**	**Predicted outcome (probability)**	**Expected outcome**	**Adverse outcome (*in vivo*)**	**Ref.**
Hexachloronaphthalene (58877-88-6)	Insulating coating and flame retardant		Positive (0.87)	Positive	Atrophy of prostate and seminal vesicles with histologic hypotrophy in castrated rats (0.3–3 mg/kg/day, 10 days, po)	Stragierowicz et al.[44]
Homosalate (118-56-9)	UV filter		Negative (0.27)	Negative	No alterations in androgen-dependent tissues (up to the limit dose of 1000 mg/kg/day)	U.S. EPA, reference number: N135-249[42]
Padimate-O (21245-02-3)	UV filter		Positive (0.70)	Positive	Decrease in glans penis, bulbocavernous muscle complex, and seminal vesicle weights in castrated rats (1000 mg/kg/day, 10 days, po)	U.S. EPA, reference number: N135-249[42]
Ensulizole (27503-81-7)	UV filter		Negative (0.001)	Negative	No alterations in androgen-dependent tissues (up to the limit dose of 1000 mg/kg/day)	U.S. EPA, reference number: N135-248[43]
Avobenzone (70356-09-1)	UV filter		Positive (0.73)	Negative	No alterations in androgen-dependent tissues (up to the limit dose of 1000 mg/kg/day)	U.S. EPA, reference number: N135-248[43]
Miltefosine (58066-85-6)	Antileishmanial drug		Positive (0.98)	Positive	Atrophy of prostate and seminal vesicles in rats (3.16–6.19 mg/kg/day, 52 weeks)	FDA, application number: 204684Orig1s000[45]
Oxyfluorfen (42874-03-3)	Herbicide		Positive (0.92)	Positive	Reduced prostate/seminal vesicle weight in castrated rats (62.5 mg/kg/day, 10 days, po)	Murr et al.[46]
Roflumilast (162401-32-3)	Anti-inflammatory drug		Positive (0.90)	Positive	Atrophy of prostate and seminal vesicles in mice, rats, hamsters, and dogs (4.0, 0.8, 4.0, and 0.6 mg/kg/day, respectively)	FDA, application number: 022522orig1s000[47]
Enzalutamide (915087-33-1)	Androgen antagonist drug		Positive (0.93)	Positive	Atrophy of prostate and seminal vesicles in rats (30 mg/kg/day, 26 weeks)	FDA, application number: 203415Orig1s000[48]
Obeticholic acid (459789-99-2)	Bile acid drug		Positive (0.70)	Positive	Reduced testes/epididymides size and prostate atrophy in rats (100–200 mg/kg/day)	FDA, application number: 207999Orig1s000[49]
O-tolidine (119-93-7)	Intermediate for dye production		Positive (0.91)	Positive	Seminal vesicle hypertrophy in immature rats (10–20 mg/kg/day, 7 days, ip)	Shangguan et al.[50]
N-Phenyl-2-naphthylamine (135-88-6)	Rubber antioxidant		Negative (0.005)	Positive	Seminal vesicle hypertrophy in immature rats (5–20 mg/kg/day, 7 days, ip)	Shangguan et al.[50]
Cylindrospermopsin (143545-90-8)	Toxin produced by freshwater cyanobacteria		Negative (0.011)	Negative	No alterations in androgen-dependent tissues in castrated rats (22.5–90 µg/kg/day)	Casas-Rodríguez et al.[51]

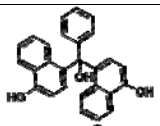
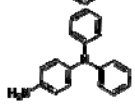
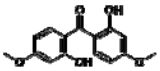
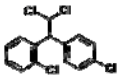
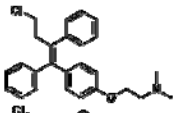
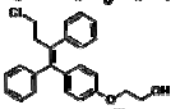
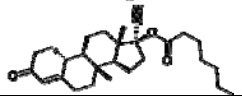
Linalool (78-70-6)	Essential oil		Negative (0.001)	Negative	No alterations in androgen-dependent tissues (100–1000 µg/kg/day)	Hareng et al.[52]
7α-Methyltestosterone (7642-58-2)	Androgen agonist candidate		Positive (0.94)	Positive	Increase in ventral prostate and seminal vesicle weights in rats (0.7–11.2 µmol/rat, 7 days, sc)	Lee et al.[53]
7α-Ethyltestosterone	Androgen agonist candidate		Positive (0.94)	Positive	Increase in ventral prostate and seminal vesicle weights in rats (0.7–11.2 µmol/rat, 7 days, sc)	Lee et al.[53]
acetyl triethyl citrate (77-89-4)	Plasticizer of cellulose resin		Positive (0.49)	Negative	No alterations in androgen-dependent tissues (20–500 µg/kg/day)	Sung et al.[54]
acetyl tributyl citrate (77-90-7)	Plasticizer found in nail polish		Negative (0.33)	Negative	No alterations in androgen-dependent tissues (20–500 µg/kg/day)	Sung et al.[54]

The application of the Tier-2 framework to the uterotrophic assay further illustrates the alignment between predicted outcomes and experimental evidence. The compound 4,4′-Butylidenebis[6-tert-butyl-3-methylphenol] (CAS no. 85-60-9), an antioxidant widely used in polyolefins and rubber applications, was correctly predicted as positive for uterotrophic activity (probability = 0.66). Experimental evidence supported this prediction, with *in vivo* uterotrophic studies reporting a significant decrease in uterine weight in mice following exposure at 100 mg/kg/day for 7 days.[55] In contrast, 2-Ethylhexyl diphenyl phosphite (CAS no. 15647-08-2), an antioxidant commonly employed in plastics and rubber manufacturing, was predicted as negative for uterotrophic activity (probability = 0.025). This prediction was consistent with *in vivo* data, as no alterations in uterine weight were detected in mice exposed to doses ranging from 15 to 1000 mg/kg/day over a 7-day period.[55] These results demonstrate that the multimodal cross-attentive GNNs extend beyond internal validation, effectively generalizing to novel compounds with organism-level endocrine effects. Such consistency with independent *in vivo* evidence underscores the reliability of the framework and highlights its potential to inform regulatory evaluation where traditional testing remains limited.

Table 2. External validation of the uterotrophic model with chemicals identified from the literature.

Compound (CAS)	**Usage**	**Structure**	**Predicted outcome (probability)**	**Expected outcome**	**Adverse outcome (*in vivo*)**	**Ref.**
1,1,1-Tris(4-hydroxyphenyl)ethane (27955-94-8)	Crosslinking agent for polymer synthesis		Positive (0.67)	Positive	Increased uterine weight in mice (300 mg/kg/day, 7 days)	Igarashi et al.[55]

4,4'-butylidenebis[6-tert-butyl-3-methylphenol] (85-60-9)	Antioxidant for polyolefins and rubber		Positive (0.66)	Positive	Decreased uterine weight in mice (100 mg/kg/day, 7 days)	Igarashi et al.[55]
2-isopropoxy-2-phenylacetophenone (6652-28-4)	Photoinitiator for UV-curable resins and coatings		Negative (0.21)	Negative	No uterine weight alterations were observed at 30–1000 mg/kg/day over 7 days	Igarashi et al.[55]
Trimethylolpropane trimethacrylate (3290-92-4)	Crosslinking agent for polymer synthesis		Negative (0.024)	Negative	No uterine weight alterations were observed at 30–300 mg/kg/day over 7 days	Igarashi et al.[55]
Triphenyl phosphate (115-86-6)	Flame retardant		Positive (0.56)	Negative	No uterine weight alterations were observed at 30–1000 mg/kg/day over 7 days	Igarashi et al.[55]
2,4-Di-tert-butylphenol (96-76-4)	UV stabilizer intermediate		Negative (0.046)	Negative	No uterine weight alterations were observed at 30–300 mg/kg/day over 7 days	Igarashi et al.[55]
Dibenzoylmethane (120-46-7)	Vinyl chloride stabilizer		Negative (0.054)	Positive	Uterine weight reduction in mice (30 mg/kg/day, 7 days, subcutaneous)	Igarashi et al.[55]
Tricresyl phosphate (78-32-0)	Flame retardant		Positive (0.63)	Positive	Uterine weight reduction in mice (30 mg/kg/day, 7 days, subcutaneous)	Igarashi et al.[55]
Triphenylsilanol (791-31-1)	Silicone polymer intermediate		Positive (0.59)	Positive	Uterine weight reduction in mice (30 mg/kg/day, 7 days, subcutaneous)	Igarashi et al.[55]
2-Ethylhexyl diphenyl phosphite (15647-08-2)	Antioxidant for plastics and rubber		Negative (0.025)	Negative	No uterine weight alterations were observed at 15–1000 mg/kg/day over 7 days	Igarashi et al.[55]
Linalool (78-70-6)	Essential oil		Negative (0.026)	Negative	No increase in estrogen-dependent tissues (100–1000 µg/kg/day)	Hareng et al.[52]
Raloxifene (198481-32-2)	Selective estrogen receptor modulator		Positive (0.64)	Positive	Increased uterine weight in immature rats (0.5–5 mg/kg over 3 days)	Komm et al.[56]
Metamizole (68-89-3)	Antipyretic drug		Negative (0.032)	Negative	No uterine weight alterations were observed at 50–200 mg/kg/day over 3 days	Passoni et al.[57]
Tetrazolium violet (1719-71-7)	Cell viability dye		Positive (0.63)	Negative	No uterine weight alterations were observed at 30 mg/kg/day over 7 days	Ohta et a.[58]
Physostigmine (57-64-7)	Antiglaucoma drug		Negative (0.02)	Negative	No uterine weight alterations were observed at 3 mg/kg/day over 7 days	Ohta et a.[58]
o-Cresolphthalein (596-27-0)	pH indicator and reagent for calcium determination		Positive (0.56)	Negative	No uterine weight alterations were observed at 1000 mg/kg/day over 7 days	Ohta et a.[58]
Permanent Orange (12236-62-3)	Synthetic azo dye for textiles and plastics		Negative (0.049)	Negative	No uterine weight alterations were observed at 300 mg/kg/day over 7 days	Ohta et a.[58]
2-[Bis(4-hydroxyphenyl)methyl] benzylalcohol (81-92-5)	Synthetic intermediate in dye		Positive (0.65)	Positive	Increased uterine weight in mice (100 mg/kg/day over 7 days)	Ohta et a.[58]
Neofuchsine (3248-91-7	Biological stain and textile dye		Positive (0.66)	Positive	Increased uterine weight in mice (300 mg/kg/day over 7 days)	Ohta et a.[58]

α-Naphtholbenzein (6948-88-5)	pH indicator dye		Positive (0.61)	Positive	Increased uterine weight in mice (300 mg/kg/day over 7 days)	Ohta et a.[58]
N,N-Difenil-p-fenilenodiamina (2350-01-8)	Antioxidant and stabilizer for rubber and lubricants		Positive (0.66)	Positive	Increased uterine weight in mice (1000 mg/kg/day over 7 days)	Ohta et a.[58]
2,2'-Dihidroxi-4,4'-dimetoxibenzofenona (131-54-4)	UV absorber and stabilizer in plastics		Positive (64)	Positive	Increased uterine weight in mice (1000 mg/kg/day over 7 days)	Ohta et a.[58]
Mitotane (53-19-0)	Antineoplastic drug		Positive (0.66)	Positive	Case report of a child with uterine enlargement (2.8 g/m²/day, 10 months), reversed by anastrozole (1 mg/day, 3 weeks)	Riedmeier et al.[59]
Toremifene (89778-26-7)	Selective estrogen receptor modulator		Positive (0.68)	Positive	Increased uterine weight in rats (1 mg/kg/day subcutaneous, 3 days); partial estrogen agonist	Carthew et al.[60]
Ospemifene (128607-22-7)	Selective estrogen receptor modulator		Positive (0.68)	Positive	Increased uterine weight in rats (0.3–10 mg/kg/day, 28 days)	Qu et al.[61]
Norethisterone enantate (3836-23-5)	Contraceptive drug		Positive (0.68)	Positive	Increased uterine weight in rats (2–4 mg/kg/day, 10 days)	Srivastava (1981)[62]

2.5. Tier-2 explainability

Besides assessing the model's performance, it is often beneficial to examine the "black box" of the trained model to gain a deeper understanding of which molecular substructures and pathway-level assays most strongly drive *in vivo* endocrine disruption. To enable explainability, the present framework incorporates two original strategies designed in this study: cross-attention mapping and counterfactual perturbations. Cross-attention analysis was performed bidirectionally, mapping molecular substructures to pathway assays (M2P) and, conversely, linking pathway nodes back to their most influential molecular features (P2M). This dual perspective enabled the identification of explicit connections between atomic features and AR- and ER-mediated assay nodes, highlighting how chemical motifs align with pathway-level perturbations. Complementing this mechanistic mapping, counterfactual perturbations were employed to generate alternative scenarios, enabling the evaluation of how molecular modifications or pathway disruptions may alter downstream outcomes.

To illustrate how the Tier-2 framework provides mechanistic interpretability for endocrine disruption predictions, we first examined two representative compounds (Figure 5): hexachloronaphthalene (CAS no. 58877-88-6), classified as positive in the Hershberger assay, and 1,1,1-tris(4-hydroxyphenyl)ethane (CAS no. 27955-94-8), identified as positive in the uterotrophic assay.

The model's classification of hexachloronaphthalene as a positive compound in the Hershberger assay is attributed to its function as an AR antagonist (probability = 0.87). The primary evidence for this was established through a cross-attention analysis, which elucidated the mechanistic underpinnings of the prediction (Figure 5a). This analysis identified strong, bidirectional connections between the molecule's chlorinated aromatic motifs and specific nodes within the AR-mediated pathway. Such a connection confirms a robust structure-activity relationship, indicating that these structural features are integral to the observed toxicological activity. This conclusion was further substantiated by counterfactual perturbations, which demonstrated that computational modification of the chlorinated ring system would negate the predicted outcome, thereby confirming the structural feature's causal role in the compound's antagonistic profile.

The model's reasoning is further clarified by tracing the compound's influence along the AOP. The analysis assigned a negative influence (indicated by a blue color) to early-stage events, including the MIE of receptor binding and subsequent key events, such as coactivator recruitment (KE1) and transcription factor regulation (KE3). This pattern suggests the compound does not operate via a classical agonistic mechanism that would activate the pathway from its inception. In contrast, the model identified a strong positive influence (indicated by a red color) concentrated in downstream, antagonistic assays. These correspond to KE4, which involves the regulation of androgen-dependent gene expression, and KE5, which pertains to cell proliferation. The model thus inferred an antagonistic mechanism from this distinct AOP signature, characterized by the absence of initial pathway activation coupled with a significant blockage of terminal biological functions.

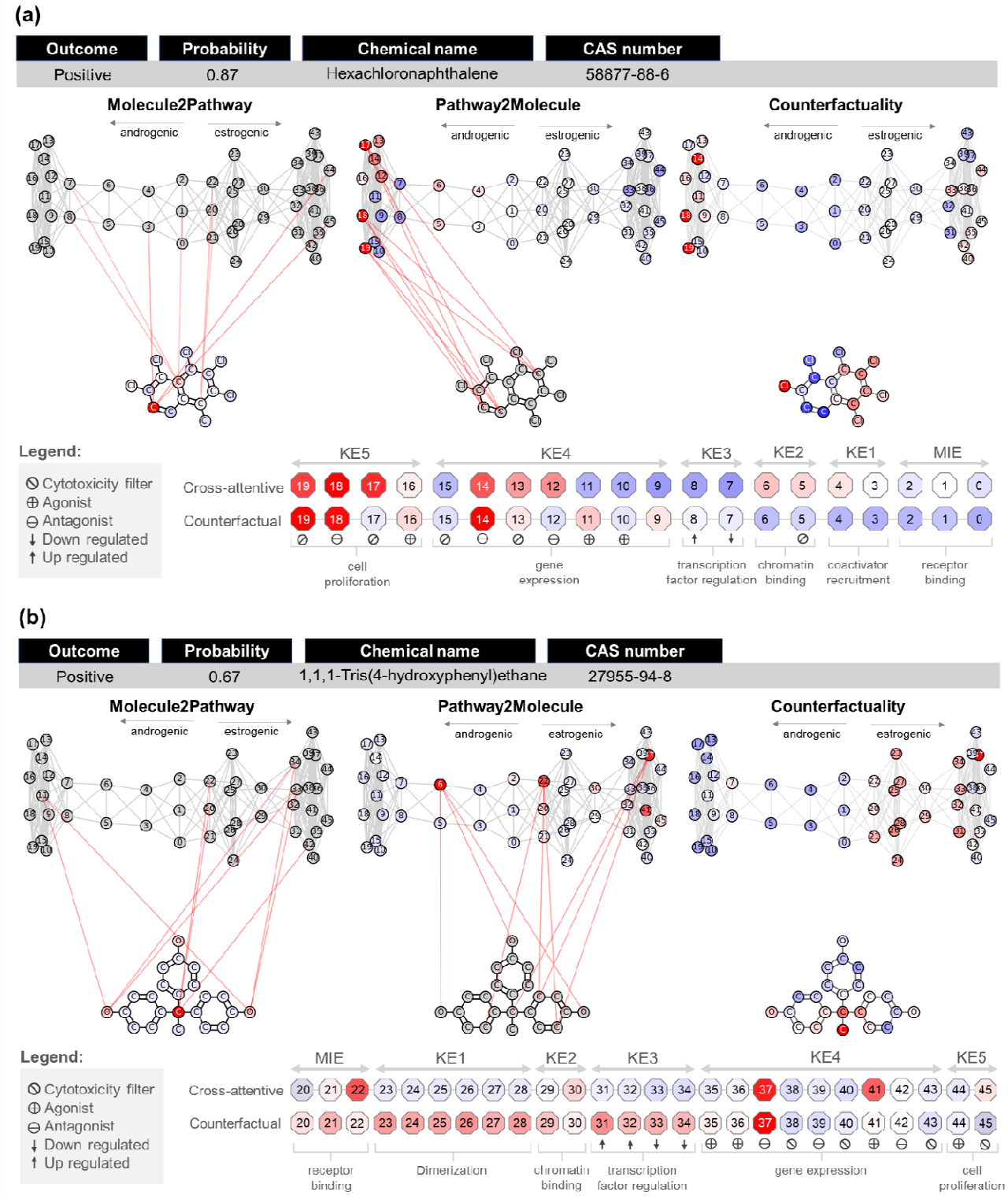


Figure 5. Integrative cross-attention (non-causal) and counterfactual analyses illustrating the mechanistic interpretability of Tier-2 predictions for (**a**) hexachloronaphthalene (Hershberger positive) and (**b**) 1,1,1-tris(4-hydroxyphenyl)ethane (uterotrophic positive).

The mechanistic insights derived from the model show a high degree of concordance with experimental *in vivo* data. Scientific literature reports that exposure to hexachloronaphthalene results in prostate and seminal vesicle atrophy in rats, a hallmark of AR antagonism. The model's identification of pathway disruption at KE4 and KE5 provides the direct mechanistic explanation for this adverse outcome. Androgen-dependent tissues require continuous signaling through the AR to maintain cellular homeostasis and mass. By blocking this pathway, hexachloronaphthalene inhibits the transcription of essential androgen-dependent genes (disruption of KE4). The subsequent failure of cells to proliferate and survive (disruption of KE5) leads to a net loss of tissue mass, which is physically observed as atrophy. This causal chain, validated by the findings of Stragierowicz et al.,[44] directly links the model's AOP-based prediction to the empirically observed *in vivo* endpoint, supporting the compound's classification as positive in the Hershberger model.

The model classifies 1,1,1-Tris(4-hydroxyphenyl)ethane as a positive compound for uterotrophic activity, providing a mechanistic rationale for its 0.67 prediction probability. Cross-attention analysis identified the molecule's aromatic hydroxyl (-OH) groups as the primary structural determinants of its bioactivity. Strong, bidirectional connections were established between these functional moieties and nodes within the estrogen receptor (ER)-mediated pathway (Figure 5b). This indicates that the model not only found a correlation but also identified these hydroxyl groups as central to the compound's mechanism of action, directly linking chemical structure to predicted toxicological function.

An examination of the model's reasoning along the AOP reveals a nuanced and complex mechanistic signature. The prediction was not driven by a simple agonistic signal; instead, the model weighed conflicting evidence. A positive influence (red) was attributed to antagonistic assays at the MIE, KE2 (chromatin binding), and a mixed agonist/antagonist profile at KE4 (gene expression). This was contrasted with a negative influence (blue) at KE1 (dimerization) and KE5 (cell proliferation). This complex pattern aligns with the "dual behavior" of certain estrogenic compounds, which can exhibit both agonistic and antagonistic features depending on the toxicological context, similar to Selective Estrogen Receptor Modulators (SERMs).[63] Counterfactual analysis further layered this complexity, showing that KE3 (transcription-factor regulation) and other KE4 assays also contributed significantly to the positive prediction.

A compelling validation of the model's assessment is found in its alignment with the definitive *in vivo* endpoint. Experimental evidence from Igarashi et al.[55] confirms that the administration of 1,1,1-Tris(4-hydroxyphenyl)ethane induces a significant increase in uterine weight in ovariectomized

mice, the classical uterotrophic response. This demonstrates that, despite the partial antagonistic signals detected across the AOP, the compound's net physiological effect in uterine tissue is predominantly agonistic, resulting in cellular proliferation and tissue growth. The model's ability to process these seemingly contradictory signals and still forecast the correct organism-level outcome showcases its capacity to learn complex, non-linear biological signatures and accurately integrate them into a final, validated prediction.

3. CONCLUSIONS

This work introduces an innovative multimodal cross-attentive framework that fuses chemical graph information with AOP-anchored, assay-level signals to predict organism-level endocrine outcomes. The multi-tier design, composed of multitask Tier-1 models that learn AR/ER MIEs and KEs, followed by Tier-2 cross-attentive models, achieved high and consistent predictive performance for both Hershberger and uterotrophic endpoints. It outperformed single-source baselines in ablation studies and generalized well in external validations. To render predictions interpretable, cross-attention analysis is performed bidirectionally, mapping molecular substructures to pathway assays (M2P) and, conversely, linking pathway nodes back to their most influential molecular features (P2M). This dual perspective establishes explicit correspondences between atomic features and AR- and ER-mediated assay nodes across the AOP (MIE→KE5), clarifying how chemical motifs align with pathway-level perturbations. Coupled with counterfactual perturbations that quantify the impact of targeted molecular edits or pathway shocks on the predicted outcome, the framework identifies the assays that most determine each *in vivo* decision and the structural elements that drive those signals. In summary, this computational approach represents an early IATA-compatible tool for screening and prioritization, helping identify substances with the highest expected decision impact for further experimental evaluation. Its use can substantially reduce redundant animal testing and optimize resource allocation, while still acknowledging key limitations, such as the chemical space covered, the lack of metabolic competence in many *in vitro* sources, potential species- and strain-specific differences, untested dose–response regimens, and the inability to fully account for mixture effects. By making such constraints explicit, the framework facilitates evidence-based integration with regulatory testing strategies, thereby contributing to a more efficient and mechanistically informed risk assessment process.

4. METHODS

4.1. Data collection and curation

A dataset was compiled from the U.S. EPA ToxCast/Tox21 program (invitroDB v4.1), encompassing 46 *in vitro* assays relevant to AR- and ER-mediated AOPs. For the AR pathway, 18 assays were included, covering receptor binding, co-regulator recruitment, transcription factor activity, gene expression, and cell proliferation. For the ER pathway, 22 assays were considered, including receptor binding, dimerization and co-regulator recruitment, chromatin binding, transcription factor activity, gene expression, and cell proliferation. Cytotoxicity assays were also incorporated to interpret antagonistic activity. In parallel, *in vivo* data were obtained from rodent uterotrophic[39] and Hershberger[38] assays. To mitigate data scarcity, datasets were augmented with additional inactive chemicals from ToxCast/Tox21 and the U.S. EPA EDSP Tier-1 assessments. Subsequently, chemical structures in SMILES format and their bioactivity records were curated according to established best practices.[64,65] The process involved standardization of functional groups and aromatic systems and removal of salts, mixtures, polymers, and organometallics. Duplicate entries with conflicting activity outcomes were excluded, while concordant duplicates were merged into a single representative record. Subsequently, the datasets were randomly split into training, validation, and test sets.

4.2. Tier-1 models

Tier-1 models were developed to predict the outcomes of 46 *in vitro* assays encompassing AR- and ER-mediated MIEs and KEs, using molecular graphs as feature representations.

4.2.1. Molecular graphs

Molecular graphs were constructed from SMILES strings using RDKit v.2024.4.5, representing atoms as nodes and bonds as edges. Atom-level descriptors comprised one-hot encodings of atomic number, degree, formal charge, hybridization, and aromaticity. Bond features were encoded through one-hot representations of bond type, conjugation, ring membership, stereochemistry, and chirality.

4.2.2. Architecture

The MPNN, GAT, and GIN architectures were refined through systematic incorporation of skip connections, virtual nodes, and Jumping Knowledge (JK) mechanisms to enhance chemical semantics. Furthermore, AttentiveFP architecture was employed without modification, as it inherently incorporates multiscale mechanisms analogous to those introduced in the optimized

models. Skip connections were incorporated to mitigate over-smoothing and gradient degradation during deep message-passing as follows:

$$h^{(\ell)} \leftarrow h^{(\ell)} + h^{(\ell-1)}$$

where $h^{(\ell-1)}$, $h^{(\ell)} \in \mathbb{R}^d$ are input and output features at layer ℓ. This additive transform acts as an identity initialization, preserving the gradient flow.

The virtual node mechanism was incorporated into each molecular graph to enrich chemical semantics by forging long-range connections among pharmacophoric and structural features. This node is initialized with a learnable embedding by $h_{virt}^{(0)} \in \mathbb{R}^d$ and interacts with all nodes by contributing additively to their representations at each message-passing layer. For a given graph $G = (V, E)$, the representation of each atom $v \in V$ at layer ℓ is modified as follows:

$$h_v^{(\ell)} \leftarrow h_v^{(\ell)} + h_{virt}^{(\ell)}$$

where $h_v^{(\ell)}$ is the atom's representation at layer ℓ, and $h_{virt}^{(\ell)}$ is the virtual node's representation at the same layer. After message-passing, the virtual node embedding is updated by aggregating the current node states across graph:

$h_{pool}^{(\ell)} = \sum_{v \in V} h_v^{(\ell)}$where $h_{pool}^{(\ell)}$ is the graph-level summary at layer ℓ, computed as permutation-invariant sum over all atomic representations $h_v^{(\ell)}$. The pooled representation is then transformed through a learnable nonlinear function $MLP^{(\ell)}$, typically consisting of two linear layers with intermediate activation and normalization, to produce a virtual node embedding as follows:

$$\Delta h_{virt}^{(\ell+1)} = MLP^{(\ell)}\,(h_{pool}^{(\ell)}) \quad (4)$$

Finally, the virtual node embedding is updated by residual addition:

$$h_{virt}^{(\ell+1)} = h_{virt}^{(\ell)} + \Delta h_{virt}^{(\ell+1)}$$

where the virtual node embedding from the previous layer $h_{virt}^{(\ell)} \in \mathbb{R}^d$ is incremented by the newly computed update $\Delta h_{virt}^{(\ell+1)} \in \mathbb{R}^d$.

The JK mechanism was incorporated into architectures to address this limitation by allowing the model to adaptively aggregate information from multiple message-passing depths, effectively capturing signals at varying neighborhood ranges. Formally, let $\mathrm{H}^{(0)}$, $\mathrm{H}^{(1)}, \ldots,$ $\mathrm{H}^{(L)}$ denote the sequence of node embeddings obtained at each of the L message-passing layers. JK aggregation builds a unified node representation through feature concatenation:

$$\mathrm{H}^{\mathrm{JK}} = \mathrm{concat}\,[\mathrm{H}^{(0)}, \mathrm{H}^{(1)}, \ldots, \mathrm{H}^{(L)}]$$

4.2.3. Homoscedastic uncertainty

A homoscedastic uncertainty weighting framework,[66] was employed to balance task contributions during multitask learning by accounting for task-dependent variance. For each task t, a trainable parameter $a_t = log\sigma_t^2$ defined as the logarithm of the task-dependent variance was introduced to adaptively scale the classification loss as follows:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sum_{i=1}^{N} m_{i.t}}\sum_{i=1}^{N} m_{i.t}\left(e^{-a_t}\ell_{i,t} + a_t\right)$$

where σ_t^2 represents the homoscedastic uncertainty associated with task t, T is the number of tasks and N the number of samples, and individual task loss $\ell_{i,t}$ corresponds to the binary cross-entropy. The binary mask $m_{i.t} \in \{0,1\}$ indicates whether a valid label is available for sample i in task t, ensuring that only valid entries contribute to the loss. The exponential term e^{-a_t} adaptively downweights tasks with higher uncertainty, while the additive a_t term penalizes overly large uncertainty estimates, preventing trivial solutions.

4.2.4. Probabilistic calibration

Probabilistic calibration was applied in class-imbalanced settings using a threshold-moving strategy[37] to adjust the decision boundaries based on raw predicted probabilities. Thresholds were evaluated individually for each MTL task across a 0.0–1.0 range with 0.01 increments on the training and validation sets, and optimal values were defined by the highest G-mean as follows:

$$\tau_t^* = \underset{\tau\in\{0.00,0.01,\ldots,1.00\}}{\arg max}\sqrt{recall_t(\tau)\cdot SP_t(\tau)}, \quad \hat{y}_{i,t}(\tau) = 1\left[\hat{p}_{i,t} \geq \tau\right]$$

where $recall_t(\tau)$ and $SP_t(\tau)$ denote recall and specificity, computed at threshold τ for task t. The indicator function $1\left[\hat{p}_{i,t} \geq \tau\right]$ assigns a binary label $\hat{y}_{i,t}(\tau)$ to each prediction $\hat{p}_{i,t}$, with outcome 1 if the predicted probability is greater than or equal to τ, and 0 otherwise. The calibrated thresholds τ_t^* were subsequently applied to the test set to reclassify predictions, enhancing the trade-off between Recall and SP while maintaining AUROC.

4.3. Tier-2 models

Tier-2 models were developed to predict adverse outcomes from the *in vivo* Hershberger and uterotrophic assays, integrating molecular and pathway graphs as complementary feature representations.

4.3.1. Pathway graphs

Pathway graphs were constructed from AR- and ER-mediated AOPs using hierarchical and co-occurrence information derived from the ToxCast/Tox21 dataset. Assays were grouped according to their MIE and KE classification, with rule levels assigned to reflect their position within AOPs. Nodes represented individual assays, annotated with Tier-1 logits, which are the raw outputs of the predictive model that quantify the strength and direction of the prediction. Edges were defined based on both structural hierarchy (linking events within the same pathway group and adjacent MIE/KE levels) and statistical associations derived from binarized Tier-1 model predictions. Association weights were computed using normalized pointwise mutual information (nPMI) and Yule's Q values, calculated from predicted class co-occurrence patterns, and incorporated as edge attributes. The nPMI between tasks a and b was given by:

$$nPMI(a,b) = \frac{log\left(\frac{p(a,b)}{p(a)\,p(b)}\right)}{-\log p(a,b)} \in [-1,1]$$

where $p(a,b)$ is the joint probability of both tasks being predicted as positive, and $p(a)$ and $p(b)$ are their respective marginal probabilities. In parallel, Yules's Q values derived from the odds ratio was computed as follows:

$$OR(a,b) = \frac{TP \cdot TN}{FP \cdot FN}, \qquad Q(a,b) = \frac{OR-1}{OR+1} \in [-1,1]$$

where TP, TN, FP, and FN denote the counts of true positives, true negatives, false positives and false negatives for the pair (a,b).

4.3.2. Architecture

Four multimodal GNNs were developed using MPNN, GAT, GIN, and AttentiveFP message-passing mechanisms. Each model employed parallel message-passing arms for the molecular and pathway graphs, enhanced with skip connections, virtual nodes, and JK mechanisms. The outputs of both encoders were projected into a shared latent dimension and integrated through a bidirectional cross-attention mechanism. Multi-head attention (MHA) was computed as:

$$\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_\mathrm{h})W^O$$

$$\mathrm{head}_\mathrm{i} = \mathrm{softmax}\left(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right)VW_i^V$$

where Q, K, and V represent query, key, and value projections of the input node, d_k is the key dimension, and W^O, W_i^Q, W_i^K, and W_i^V are learned projection matrices. In the M2P direction, pathway embeddings were updated as:

$$\mathrm{P}' = \mathrm{P} + \mathrm{MHA}(\mathrm{P}, \mathrm{M}, \mathrm{M}))$$

where P denotes the input pathway node embeddings and M the molecular embeddings. Conversely, in the P2M direction, molecular representations were updated as:

$$\mathrm{M}' = \mathrm{M} + \mathrm{Attention}(\mathrm{M}, \mathrm{P}', \mathrm{P}')$$

where P' are the updated pathway embeddings. Finally, the refined embeddings (P' and M') were pooled, concatenated, and passed through a multilayer perceptron to generate *in vivo* predictions

4.4. Model training and evaluation

All models were implemented in PyTorch v.2.5.1 with PyTorch Geometric v.2.6.1 and trained on an NVIDIA TITAN Xp GPU. Training was capped at 2,000 epochs using binary cross-entropy (BCE) loss, with early stopping applied when no improvement in validation loss was observed within a five-epoch adaptive window. A ReduceLROnPlateau scheduler was employed to reduce the learning rate when validation loss stagnated beyond the defined patience. To enhance efficiency, pruning techniques were integrated to halt underperforming trials. Hyperparameter tuning was performed using Optuna v.4.2.0, with 200 trials per architecture to systematically explore the parameter space. Model performance was assessed using ACC, recall, SP, and AUROC, with G-mean additionally considered to address class imbalance. Embedding vectors generated during training were retained and later projected via t-SNE and KDE to evaluate the structuring of the latent space across epochs.

4.5. Cross-attentive explainability

Cross-attentive explanations were generated by extracting bidirectional attention matrices between the molecular and pathway graphs during inference, where M2P and P2M attention weights were used to compute node-level importance scores. Formally, the importance score for a source node *i* was calculated as:

$$h_i = \frac{1}{|T(i)|} \sum_{j \in T(i)} A_{ij}$$

where A_{ij} denotes the cross-attention weight from node i to target j, and $T(i)$ is the set of valid attended targets. In the M2P direction, i corresponds to a molecular node (atom) and j to a pathway node (assay). Conversely, in the P2M direction, i denotes a pathway node and j a molecular node. Interpretability was enhanced by retaining only the top-k attention links between i and j. Finally, both graphs were jointly rendered within the same spatial frame, with bidirectional cross-attention links superimposed to highlight the most influential molecular–assay interactions.

4.6. Counterfactual explainability

Stochastic perturbations were applied to the molecular and pathway graphs using rule-based isosteric and valence-preserving substitutions, and quasi-random assay perturbations derived from low-discrepancy sequences, respectively. Subsequently, the importance score of each node i was calculated as follows:

$$h_i = \frac{1}{K}\sum_{k=1}^{K}\left(p_0 - p_i^{(k)}\right)$$

where p_0 is the predicted probability for the unperturbed graph, $p_i^{(k)}$ is the predicted probability after k-th perturbation of node i, and K is the number of perturbation samples. Graph-based clustering was then performed using connected components to group structurally connected atoms with similar contribution signals, with intra-cluster enhancement applied as:

$$C_i = \mu_c + \gamma(h_i - \mu_c)$$

where C_i is the cluster-enhanced contribution score for atom i, $\mu_c = \frac{1}{|c|}\sum_{j\in c} h_i$ is the mean contribution c, and $\gamma > 0$ is the cluster-enhancement factor. Laplacian smoothing was subsequently applied within each cluster to reduce noise while preserving local structure as follows:

$$\tilde{h}_i = \alpha C_i + (1-\alpha)\sum_{j\in N(i)}\frac{1}{d_i}C_j$$

where C_j is the cluster-enhanced score of each neighbor $j \in N(i)$, whereas $\alpha \in [0, 1]$ balances self-vs-neighbor contributions, $N(i)$ is the set of neighbors of atom i, and $d_i = |N(i)|$ its degree. Finally, the $\tilde{h}_i$ values are normalized (zero mean, unit variance) and projected onto 2D graph coordinates for visualization.

Acknowledgements

The authors thank CNPq, CAPES, and FAPEG for financial support and fellowships. This study received financial support from the FAPEG (grant # 202310267001412), CAPES (Finance Code 001), and INCT (grant #408678/2024-0). BJN is CNPq productivity fellow (grant #311100/2023-6).

Conflict of interest

RCB and ENM are co-founders of InsilicAll LLC and Predictive LLC, respectively, which develop novel alternative methods and software for toxicity prediction. All the other authors declare no conflicts.

REFERENCES

(1) Shanle, E. K.; Xu, W. Endocrine Disrupting Chemicals Targeting Estrogen Receptor Signaling: Identification and Mechanisms of Action. *Chem Res Toxicol* **2011**, *24* (1), 6–19. https://doi.org/10.1021/tx100231n.

(2) Rizzo, R.; Bortolotti, D.; Rizzo, S.; Schiuma, G. Cellular Mechanisms of Endocrine Disruption. In *Environment Impact on Reproductive Health: A Translational Approach*; Marci, R., Ed.; Springer International Publishing: Cham, 2023; pp 15–48. https://doi.org/10.1007/978-3-031-36494-5_2.

(3) Ahn, C.; Jeung, E.-B. Endocrine-Disrupting Chemicals and Disease Endpoints. *IJMS* **2023**, *24* (6), 5342. https://doi.org/10.3390/ijms24065342.

(4) Cripps, S. M.; Marshall, S. A.; Mattiske, D. M.; Ingham, R. Y.; Pask, A. J. Estrogenic Endocrine Disruptor Exposure Directly Impacts Erectile Function. *Commun Biol* **2024**, *7* (1), 403. https://doi.org/10.1038/s42003-024-06048-1.

(5) Kawa, I. A.; Akbar Masood; Fatima, Q.; Mir, S. A.; Jeelani, H.; Manzoor, S.; Rashid, F. Endocrine Disrupting Chemical Bisphenol A and Its Potential Effects on Female Health. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* **2021**, *15* (3), 803–811. https://doi.org/10.1016/j.dsx.2021.03.031.

(6) Lacouture, A.; Lafront, C.; Peillex, C.; Pelletier, M.; Audet-Walsh, É. Impacts of Endocrine-Disrupting Chemicals on Prostate Function and Cancer. *Environmental Research* **2022**, *204*, 112085. https://doi.org/10.1016/j.envres.2021.112085.

(7) Soto, A. M.; Sonnenschein, C. DDT, Endocrine Disruption and Breast Cancer. *Nat Rev Endocrinol* **2015**, *11* (9), 507–508. https://doi.org/10.1038/nrendo.2015.125.

(8) Trasande, L.; Zoeller, R. T.; Hass, U.; Kortenkamp, A.; Grandjean, P.; Myers, J. P.; DiGangi, J.; Bellanger, M.; Hauser, R.; Legler, J.; Skakkebaek, N. E.; Heindel, J. J. Estimating Burden and Disease Costs of Exposure to Endocrine-Disrupting Chemicals in the European Union. *J Clin Endocrinol Metab* **2015**, *100* (4), 1245–1255. https://doi.org/10.1210/jc.2014-4324.

(9) Manibusan, M. K.; Touart, L. W. A Comprehensive Review of Regulatory Test Methods for Endocrine Adverse Health Effects. *Crit Rev Toxicol* **2017**, *47* (6), 433–481. https://doi.org/10.1080/10408444.2016.1272095.

(10) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29* (8), 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135.

(11) Richard, A. M.; Huang, R.; Waidyanatha, S.; Shinn, P.; Collins, B. J.; Thillainadarajah, I.; Grulke, C. M.; Williams, A. J.; Lougee, R. R.; Judson, R. S.; Houck, K. A.; Shobair, M.; Yang, C.; Rathman, J. F.; Yasgar,

A.; Fitzpatrick, S. C.; Simeonov, A.; Thomas, R. S.; Crofton, K. M.; Paules, R. S.; Bucher, J. R.; Austin, C. P.; Kavlock, R. J.; Tice, R. R. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem. Res. Toxicol.* **2021**, *34* (2), 189–216. https://doi.org/10.1021/acs.chemrestox.0c00264.

(12) Jeong, J.; Kim, D.; Choi, J. Application of ToxCast/Tox21 Data for Toxicity Mechanism-Based Evaluation and Prioritization of Environmental Chemicals: Perspective and Limitations. *Toxicology in Vitro* **2022**, *84*, 105451. https://doi.org/10.1016/j.tiv.2022.105451.

(13) Kleinstreuer, N. C.; Ceger, P.; Watt, E. D.; Martin, M.; Houck, K.; Browne, P.; Thomas, R. S.; Casey, W. M.; Dix, D. J.; Allen, D.; Sakamuru, S.; Xia, M.; Huang, R.; Judson, R. Development and Validation of a Computational Model for Androgen Receptor Activity. *Chem Res Toxicol* **2017**, *30* (4), 946–964. https://doi.org/10.1021/acs.chemrestox.6b00347.

(14) Judson, R. S.; Magpantay, F. M.; Chickarmane, V.; Haskell, C.; Tania, N.; Taylor, J.; Xia, M.; Huang, R.; Rotroff, D. M.; Filer, D. L.; Houck, K. A.; Martin, M. T.; Sipes, N.; Richard, A. M.; Mansouri, K.; Setzer, R. W.; Knudsen, T. B.; Crofton, K. M.; Thomas, R. S. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol Sci* **2015**, *148* (1), 137–154. https://doi.org/10.1093/toxsci/kfv168.

(15) Thomas, R. S.; Paules, R. S.; Simeonov, A.; Fitzpatrick, S. C.; Crofton, K. M.; Casey, W. M.; Mendrick, D. L. The US Federal Tox21 Program: A Strategic and Operational Plan for Continued Leadership. *ALTEX* **2018**, *35* (2), 163–168. https://doi.org/10.14573/altex.1803011.

(16) Rotroff, D. M.; Dix, D. J.; Houck, K. A.; Knudsen, T. B.; Martin, M. T.; McLaurin, K. W.; Reif, D. M.; Crofton, K. M.; Singh, A. V.; Xia, M.; Huang, R.; Judson, R. S. Using in Vitro High Throughput Screening Assays to Identify Potential Endocrine-Disrupting Chemicals. *Environ Health Perspect* **2013**, *121* (1), 7–14. https://doi.org/10.1289/ehp.1205065.

(17) Rotroff, D. M.; Dix, D. J.; Houck, K. A.; Knudsen, T. B.; Martin, M. T.; McLaurin, K. W.; Reif, D. M.; Crofton, K. M.; Singh, A. V.; Xia, M.; Huang, R.; Judson, R. S. Using in Vitro High Throughput Screening Assays to Identify Potential Endocrine-Disrupting Chemicals. *Environmental Health Perspectives* **2013**, *121* (1), 7–14. https://doi.org/10.1289/ehp.1205065.

(18) Browne, P.; Noyes, P. D.; Casey, W. M.; Dix, D. J. Application of Adverse Outcome Pathways to U.S. EPA's Endocrine Disruptor Screening Program. *Environmental Health Perspectives* **2017**, *125* (9), 096001. https://doi.org/10.1289/EHP1304.

(19) Thomas, R. S.; Bahadori, T.; Buckley, T. J.; Cowden, J.; Deisenroth, C.; Dionisio, K. L.; Frithsen, J. B.; Grulke, C. M.; Gwinn, M. R.; Harrill, J. A.; Higuchi, M.; Houck, K. A.; Hughes, M. F.; Hunter, E. S., III; Isaacs, K. K.; Judson, R. S.; Knudsen, T. B.; Lambert, J. C.; Linnenbrink, M.; Martin, T. M.; Newton, S. R.; Padilla, S.; Patlewicz, G.; Paul-Friedman, K.; Phillips, K. A.; Richard, A. M.; Sams, R.; Shafer, T. J.; Setzer, R. W.; Shah, I.; Simmons, J. E.; Simmons, S. O.; Singh, A.; Sobus, J. R.; Strynar, M.; Swank, A.; Tornero-Valez, R.; Ulrich, E. M.; Villeneuve, D. L.; Wambaugh, J. F.; Wetmore, B. A.; Williams, A. J. The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. *Toxicol Sci* **2019**, *169* (2), 317–332. https://doi.org/10.1093/toxsci/kfz058.

(20) Chen, Y.; Cheng, F.; Sun, L.; Li, W.; Liu, G.; Tang, Y. Computational Models to Predict Endocrine-Disrupting Chemical Binding with Androgen or Oestrogen Receptors. *Ecotoxicology and Environmental Safety* **2014**, *110*, 280–287. https://doi.org/10.1016/j.ecoenv.2014.08.026.

(21) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environmental Health Perspectives* **2016**, *124* (7), 1023–1033. https://doi.org/10.1289/ehp.1510267.

(22) Mansouri, K.; Kleinstreuer, N.; Abdelaziz, A. M.; Alberga, D.; Alves, V. M.; Andersson, P. L.; Andrade, C. H.; Bai, F.; Balabin, I.; Ballabio, D.; Benfenati, E.; Bhhatarai, B.; Boyer, S.; Chen, J.; Consonni, V.; Farag, S.; Fourches, D.; García-Sosa, A. T.; Gramatica, P.; Grisoni, F.; Grulke, C. M.; Hong, H.; Horvath, D.; Hu, X.; Huang, R.; Jeliazkova, N.; Li, J.; Li, X.; Liu, H.; Manganelli, S.; Mangiatordi, G. F.; Maran, U.; Marcou, G.; Martin, T.; Muratov, E.; Nguyen, D.-T.; Nicolotti, O.; Nikolov, N. G.; Norinder, U.; Papa, E.; Petitjean, M.; Piir, G.; Pogodin, P.; Poroikov, V.; Qiao, X.; Richard, A. M.; Roncaglioni, A.; Ruiz, P.; Rupakheti, C.; Sakkiah, S.; Sangion, A.; Schramm, K.-W.; Selvaraj, C.; Shah, I.; Sild, S.; Sun, L.; Taboureau, O.; Tang, Y.; Tetko, I. V.; Todeschini, R.; Tong, W.; Trisciuzzi, D.; Tropsha, A.; Van Den Driessche, G.; Varnek, A.; Wang, Z.; Wedebye, E. B.; Williams, A. J.; Xie, H.; Zakharov, A. V.; Zheng, Z.; Judson, R. S. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives* **2020**, *128* (2), 027002. https://doi.org/10.1289/EHP5580.

(23) Schaduangrat, N.; Anuwongcharoen, N.; Charoenkwan, P.; Shoombuatong, W. DeepAR: A Novel Deep Learning-Based Hybrid Framework for the Interpretable Prediction of Androgen Receptor Antagonists. *Journal of Cheminformatics* **2023**, *15* (1), 50. https://doi.org/10.1186/s13321-023-00721-z.

(24) Tan, H.; Wu, J.; Zhang, R.; Zhang, C.; Li, W.; Chen, Q.; Zhang, X.; Yu, H.; Shi, W. Development, Validation, and Application of a Human Reproductive Toxicity Prediction Model Based on Adverse Outcome Pathway. *Environ. Sci. Technol.* **2022**, *56* (17), 12391–12403. https://doi.org/10.1021/acs.est.2c02242.

(25) Piir, G.; Sild, S.; Maran, U. Interpretable Machine Learning for the Identification of Estrogen Receptor Agonists, Antagonists, and Binders. *Chemosphere* **2024**, *347*, 140671. https://doi.org/10.1016/j.chemosphere.2023.140671.

(26) Xin, L.; Liu, S.; Shi, W.; Ying, G.-G.; Hui, X.; Chen, C.-E. Knowledge-Based Machine Learning for Predicting and Understanding the Androgen Receptor (AR)-Mediated Reproductive Toxicity in Zebrafish. *Environment International* **2024**, *191*, 108995. https://doi.org/10.1016/j.envint.2024.108995.

(27) Piir, G.; Sild, S.; Maran, U. Interpretable Machine Learning for the Identification of Estrogen Receptor Agonists, Antagonists, and Binders. *Chemosphere* **2024**, *347*, 140671. https://doi.org/10.1016/j.chemosphere.2023.140671.

(28) Moné, M. J.; Pallocca, G.; Escher, S. E.; Exner, T.; Herzler, M.; Bennekou, S. H.; Kamp, H.; Kroese, E. D.; Leist, M.; Steger-Hartmann, T.; van de Water, B. Setting the Stage for Next-Generation Risk Assessment with Non-Animal Approaches: The EU-ToxRisk Project Experience. *Arch Toxicol* **2020**, *94* (10), 3581–3592. https://doi.org/10.1007/s00204-020-02866-4.

(29) Ball, N.; Bars, R.; Botham, P. A.; Cuciureanu, A.; Cronin, M. T. D.; Doe, J. E.; Dudzina, T.; Gant, T. W.; Leist, M.; van Ravenzwaay, B. A Framework for Chemical Safety Assessment Incorporating New Approach Methodologies within REACH. *Arch Toxicol* **2022**, *96* (3), 743–766. https://doi.org/10.1007/s00204-021-03215-9.

(30) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph Neural Networks for Materials Science and Chemistry. *Communications Materials* **2022**, *3* (1), 93. https://doi.org/10.1038/s43246-022-00315-6.

(31) Corso, G.; Stark, H.; Jegelka, S.; Jaakkola, T.; Barzilay, R. Graph Neural Networks. *Nature Reviews Methods Primers* **2024**, *4* (1), 17. https://doi.org/10.1038/s43586-024-00294-7.

(32) Soulios, K.; Scheibe, P.; Bernt, M.; Hackermüller, J.; Schor, J. deepFPlearn+: Enhancing Toxicity Prediction across the Chemical Universe Using Graph Neural Networks. *Bioinformatics* **2023**, *39* (12), btad713. https://doi.org/10.1093/bioinformatics/btad713.

(33) Torres, L. H. M.; Arrais, J. P.; Ribeiro, B. Combining Graph Neural Networks and Transformers for Few-Shot Nuclear Receptor Binding Activity Prediction. *J Cheminform* **2024**, *16* (1), 109. https://doi.org/10.1186/s13321-024-00902-4.

(34) Ren, Y.; Wu, X.; Xi, Z.; Tang, K. H. D.; Zeng, X.; Li, R.; Pan, J. Novel Graph Neural Network Reveals Binding Mechanisms and Environmental Risks of PAHs Interaction with Estrogen Receptor B. *Environmental Pollution* **2025**, *384*, 127011. https://doi.org/10.1016/j.envpol.2025.127011.
(35) Escher, B. I.; Henneberger, L.; König, M.; Schlichting, R.; Fischer, F. C. Cytotoxicity Burst? Differentiating Specific from Nonspecific Effects in Tox21 in Vitro Reporter Gene Assays. *Environ Health Perspect* **2020**, *128* (7), 077007. https://doi.org/10.1289/EHP6664.
(36) Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S. Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. *Toxicol Sci* **2016**, *152* (2), 323–339. https://doi.org/10.1093/toxsci/kfw092.
(37) Moreira-Filho, J. T.; Braga, R. C.; Lemos, J. M.; Alves, V. M.; Borba, J. V. V. B.; Costa, W. S.; Kleinstreuer, N.; Muratov, E. N.; Andrade, C. H.; Neves, B. J. BeeToxAI: An Artificial Intelligence-Based Web App to Assess Acute Toxicity of Chemicals to Honey Bees. *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100013. https://doi.org/10.1016/j.ailsci.2021.100013.
(38) Browne, P.; Kleinstreuer, N. C.; Ceger, P.; Deisenroth, C.; Baker, N.; Markey, K.; Thomas, R. S.; Judson, R. J.; Casey, W. Development of a Curated Hershberger Database. *Reprod Toxicol* **2018**, *81*, 259–271. https://doi.org/10.1016/j.reprotox.2018.08.016.
(39) Kleinstreuer, N. C.; Ceger, P. C.; Allen, D. G.; Strickland, J.; Chang, X.; Hamm, J. T.; Casey, W. M. A Curated Database of Rodent Uterotrophic Bioactivity. *Environ Health Perspect* **2016**, *124* (5), 556–562. https://doi.org/10.1289/ehp.1510183.
(40) Conolly, R. B.; Ankley, G. T.; Cheng, W.; Mayo, M. L.; Miller, D. H.; Perkins, E. J.; Villeneuve, D. L.; Watanabe, K. H. Quantitative Adverse Outcome Pathways and Their Application to Predictive Toxicology. *Environ. Sci. Technol.* **2017**, *51* (8), 4661–4672. https://doi.org/10.1021/acs.est.6b06230.
(41) Perkins, E. J.; Ashauer, R.; Burgoon, L.; Conolly, R.; Landesmann, B.; Mackay, C.; Murphy, C. A.; Pollesch, N.; Wheeler, J. R.; Zupanic, A.; Scholz, S. Building and Applying Quantitative Adverse Outcome Pathway Models for Chemical Hazard and Risk Assessment. *Environ Toxicol Chem* **2019**, *38* (9), 1850–1865. https://doi.org/10.1002/etc.4505.
(42) U.S. Environmental Protection Agency. Hershberger Bioassay for Padimate-O and Homosalate – Endocrine Disruptor Screening Program (EDSP) Tier 1 Screening., 2012. https://ntp.niehs.nih.gov/sites/default/files/2024-04/uv3_hershberger_redacted_disclaimer.pdf (accessed 2025-08-25).
(43) U.S. Environmental Protection Agency. The Hershberger Bioassay for Ensulizole and Avobenzone – Endocrine Disruptor Screening Program (EDSP) Tier 1 Screening., 2012. https://ntp.niehs.nih.gov/sites/default/files/2024-04/uv2_hershberger_redacted_disclaimer.pdf (accessed 2025-08-25).
(44) Stragierowicz, J.; Nasiadek, M.; Stasikowska-Kanicka, O.; Kolasa, A.; Kilanowicz, A. An Assessment of the (Anti)Androgenic Properties of Hexachloronaphthalene (HxCN) in Male Rats. *Chemosphere* **2024**, *359*, 142373. https://doi.org/10.1016/j.chemosphere.2024.142373.
(45) U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Application Number: 204684Orig1s000, 2013. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/204684Orig1s000SumR.pdf (accessed 2025-08-25).
(46) Murr, A. S.; Buckalew, A. R.; Devane, G.; Bailey, J. R.; Ford, J. L.; Gray, L. E.; Stoker, T. E. Peripubertal Exposure to Oxyfluorfen, a Diphenyl Herbicide, Delays Pubertal Development in the Male Rat by Antagonizing Androgen Receptor Activity. *Toxicol Sci* **2025**, *203* (2), 206–215. https://doi.org/10.1093/toxsci/kfae144.
(47) U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Application Number: 022522Orig1s000, 2010.

https://www.accessdata.fda.gov/drugsatfda_docs/nda/2011/022522orig1s000pharmr.pdf (accessed 2025-08-25).
(48) U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Application Number: 203415Orig1s000, 2012. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2012/203415Orig1s000PharmR.pdf (accessed 2025-08-25).
(49) U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Application Number: 207999Orig1s000, 2016. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2016/207999Orig1s000PharmR.pdf (accessed 2025-08-25).
(50) Shangguan, Z.; Yang, K.; Pan, Q.; Hu, H.; Wang, P. Dye Intermediates N-Phenyl-2-Naphthylamine and o-Tolidine Are Novel Environmental Androgens with Reproductive Toxicity in Male Rats. *Reproductive Toxicology* **2025**, *136*, 108955. https://doi.org/10.1016/j.reprotox.2025.108955.
(51) Casas-Rodríguez, A.; Ayala-Soldado, N.; Cameán, A. M.; Jos, A. Evaluation of Potential Androgenic and Antiandrogenic Activities of Pure Cyanotoxins (Microcystin-LR and Cylindrospermopsin) Using the Hershberger Bioassay (OECD TG 441). *Ecotoxicology and Environmental Safety* **2025**, *302*, 118780. https://doi.org/10.1016/j.ecoenv.2025.118780.
(52) Hareng, L.; Kolle, S. N.; Gomes, C.; Schneider, S.; Wahl, M. Critical Assessment of the Endocrine Potential of Linalool and Linalyl Acetate: Proactive Testing Strategy Assessing Estrogenic and Androgenic Activity of Lavender Oil Main Components. *Arch Toxicol* **2024**, *98* (1), 347–361. https://doi.org/10.1007/s00204-023-03623-z.
(53) Lee, M. S.; Bunin, D. I.; Furimsky, A. M.; Nguyen, D.; Parman, T.; Kim, K.; Rausch, L.; Lin, M.-T.; Gupta, P.; Brown, J. E.; Kroopnick, J. M.; Blithe, D. L. Novel Progestogenic Androgens for Male Contraception: Design, Synthesis, and Activity of C7 α-Substituted Testosterone†. *Biol Reprod* **2023**, *109* (6), 851–863. https://doi.org/10.1093/biolre/ioad111.
(54) Sung, C. R.; Kang, H. G.; Hong, J. Y.; Kwack, S. J. Citrate Ester Substitutes for Di-2-Ethylhexyl Phthalate: In Vivo Reproductive and in Vitro Cytotoxicity Assessments. *J Toxicol Environ Health A* **2020**, *83* (17–18), 589–595. https://doi.org/10.1080/15287394.2020.1798832.
(55) Igarashi, T.; Yokota, S.; Aida, A.; Nishimura, T.; Kitajima, S. *In Vivo* Screening Evaluation of 12 Chemicals as Candidate Endocrine Disruptors Using Ovariectomized Mouse Uterotrophic Bioassay. *Regulatory Toxicology and Pharmacology* **2025**, *162*, 105900. https://doi.org/10.1016/j.yrtph.2025.105900.
(56) Komm, B. S.; Kharode, Y. P.; Bodine, P. V. N.; Harris, H. A.; Miller, C. P.; Lyttle, C. R. Bazedoxifene Acetate: A Selective Estrogen Receptor Modulator with Improved Selectivity. *Endocrinology* **2005**, *146* (9), 3999–4008. https://doi.org/10.1210/en.2005-0030.
(57) Passoni, M. T.; Palu, G.; Grechi, N.; da Silva Amaral, B. A.; Gomes, C.; Rülker, C.; van Ravenzwaay, B.; Martino-Andrade, A. J. Uterotrophic and *in Vitro* Screening for (Anti)Estrogenic Activity of Dipyrone. *Toxicology Letters* **2021**, *352*, 1–8. https://doi.org/10.1016/j.toxlet.2021.09.004.
(58) Ohta, R.; Takagi, A.; Ohmukai, H.; Marumo, H.; Ono, A.; Matsushima, Y.; Inoue, T.; Ono, H.; Kanno, J. Ovariectomized Mouse Uterotrophic Assay of 36 Chemicals. *J. Toxicol. Sci.* **2012**, *37* (5), 879–889. https://doi.org/10.2131/jts.37.879.
(59) Riedmeier, M.; Antonini, S.; Benoit, C.; Deal, C. L.; Martin, F.; Figueiredo, B. C.; Nazli Gonc, E.; Härtel, C.; Idkowiak, J.; Kurlbaum, M.; de Krijger, R.; Ribeiro, R. C.; del Rivero, J.; Schlegel, P.-G.; Thompson, L. D. R.; Yalcin, B.; Wiegering, V. Isosexual Precocious Pseudopuberty during Mitotane Treatment in a Child with Adrenocortical Carcinoma: A Case Report. *Pediatric Hematology Oncology Journal* **2024**, *9* (2), 74–77. https://doi.org/10.1016/j.phoj.2024.03.005.
(60) Carthew, P.; Edwards, R. E.; Nolan, B. M.; Tucker, M. J.; Smith, L. L. Compartmentalized Uterotrophic Effects of Tamoxifen, Toremifene, and Estradiol in the Ovariectomized Wistar (Han) Rat. *Toxicol Sci* **1999**, *48* (2), 197–205. https://doi.org/10.1093/toxsci/48.2.197.

(61) Qu, Q.; Zheng, H.; Dahllund, J.; Laine, A.; Cockcroft, N.; Peng, Z.; Koskinen, M.; Hemminki, K.; Kangas, L.; Väänänen, K.; Härkönen, P. Selective Estrogenic Effects of a Novel Triphenylethylene Compound, FC1271a, on Bone, Cholesterol Level, and Reproductive Tissues in Intact and Ovariectomized Rats1. *Endocrinology* **2000**, *141* (2), 809–820. https://doi.org/10.1210/endo.141.2.7342.
(62) Srivastava, U. K. Effect of Intramuscular Norethisterone Enanthate on Compensatory Ovarian Hypertrophy in the Rat. *Indian J Physiol Pharmacol* **1981**, *25* (2), 151–157.
(63) Wu, X.; Yang, X.; Tian, Y.; Xu, P.; Yue, H.; Sang, N. Bisphenol B and Bisphenol AF Exposure Enhances Uterine Diseases Risks in Mouse. *Environment International* **2023**, *173*, 107858. https://doi.org/10.1016/j.envint.2023.107858.
(64) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.
(65) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *Journal of Chemical Information and Modeling* **2016**, *56* (7), 1243–1252. https://doi.org/10.1021/acs.jcim.6b00129.
(66) Cipolla, R.; Gal, Y.; Kendall, A. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2018; pp 7482–7491. https://doi.org/10.1109/CVPR.2018.00781.