Compound-Conditioned Bioactivity Graphs: Structuring Intermediate Supervision for Low-Data Toxicity Prediction

Anonymous Authors

Abstract

Developmental toxicity is difficult to predict from molecular structure alone because adverse phenotypes emerge through intermediate biological responses, while organism-level labels remain sparse, imbalanced, and incomplete. Large-scale *in vitro* bioactivity offers transferable supervision, but predicted assay profiles are typically treated as flat vectors that discard relationships among assays. We introduce compound-conditioned bioactivity graphs, in which predictions from a frozen multitask model initialize a shared, mechanistically constrained assay topology and learned node and edge gates adapt both assays and assay–assay interactions to each chemical. Tier I pretrains a molecular graph neural network across 457 ToxCast assays. Tier II integrates the conditioned assay graph with molecular and physicochemical representations using cross-modal contrastive alignment, masked multi-label representation learning, and uncertainty-weighted multitask prediction. Across five matched held-out splits on 1,415 compounds and six zebrafish developmental phenotypes, the full model achieves 0.786 ± 0.020 AUROC, compared with 0.723 ± 0.034 for flat-logit transfer, 0.718 ± 0.040 for molecular structure alone, 0.730 ± 0.016 for a feature-matched random forest, and 0.751 ± 0.024 without physicochemical descriptors. A static assay graph reaches 0.740 ± 0.025, while node-only and edge-only gating reach 0.729 ± 0.057 and 0.722 ± 0.036, respectively, indicating that both adaptive mechanisms are jointly required. Gains occur across all six phenotypes, and qualitative analyses of training, validation, and held-out compounds reveal heterogeneous assay-utilization patterns over the shared topology. These results show that the organization and compound-specific adaptation of intermediate supervision determine how effectively bioactivity transfers to low-data developmental-toxicity prediction.

TL;DR: Predicted bioactivity transfers better when organized on a mechanistically constrained assay graph whose node and edge importance adapts to each compound.

Keywords: Multimodal learning; graph neural networks; contrastive learning; transfer learning; multitask learning; computational toxicology; developmental toxicity.

1 Introduction

Tens of thousands of chemicals are in commercial use, yet *in vivo* developmental-toxicity data are unavailable for most of them. Experimental testing cannot close this gap at the required scale because of cost, throughput, and animal-use constraints. Computational developmental toxicology therefore predicts adverse phenotypes from chemical structure, but structure-only models compress the cascade from molecular perturbation through signaling and cellular responses to disrupted development into a single mapping. Omitting these intermediate biological states may limit generalization when phenotype supervision is sparse.

High-throughput programs such as ToxCast and Tox21 provide intermediate biological signals across hundreds of *in vitro* assays [Richard et al., 2016, Tice et al., 2013]. Combining bioactivity with chemical descriptors improves prediction of *in vivo* toxicity [Liu et al., 2015, 2017, Garcia de Lomana et al., 2021]; predicted assay profiles further extend this strategy to chemicals lacking experimental

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

coverage [Garcia de Lomana et al., 2021, Duran-Frigola et al., 2020, Bertoni et al., 2021]. These assay responses provide a potential bridge between chemical structure and downstream developmental outcomes.

Existing approaches generally encode bioactivity as a flat vector, treating assays as independent coordinates despite relationships through target families, signaling cascades, co-activation, and downstream cellular responses. This discards structure that may be especially important for developmental toxicity, where phenotypes emerge from coordinated perturbations across multiple mechanisms, and requires the downstream model to rediscover those relationships from limited labels.

We instead represent predicted bioactivity as a state over an assay graph. invitroDB annotations define an AOP-inspired upstream-to-downstream hierarchy [Ankley et al., 2010]; mechanistically plausible relationships form the candidate topology, and co-activation statistics provide edge attributes. The topology is shared, but its effective realization is compound specific: Tier-I logits initialize assay nodes, while learned node and edge gates adapt assays and assay–assay interactions to each chemical. This conditioning allows chemicals to occupy the same biological network while emphasizing different mechanistic routes [Perez et al., 2018].

The conditioned graph is integrated with molecular and physicochemical representations in a multimodal developmental-toxicity predictor. Cross-modal contrastive learning aligns the three views [van den Oord et al., 2018, Radford et al., 2021], while a masked multi-label objective uses only jointly observed developmental endpoints and uncertainty weighting balances prediction tasks [Kendall et al., 2018]. Missing compound–endpoint observations are masked rather than treated as negatives.

We evaluate zebrafish developmental toxicity [Truong et al., 2014], transferring supervision from 9,020 chemicals and 457 ToxCast assays to 1,415 chemicals with six partially observed phenotype groups. Across five matched held-out splits, the full model achieves 0.786 ± 0.020 AUROC, outperforming molecular-only prediction (0.718 ± 0.040), a feature-matched random forest (0.730 ± 0.016), and flat-logit transfer through an MLP (0.723 ± 0.034), with gains across all six phenotypes. A static assay graph reaches 0.740 ± 0.025, while node-only and edge-only gating reach 0.729 ± 0.057 and 0.722 ± 0.036, respectively. Qualitative inspection across training, validation, and held-out compounds further reveals distinct assay-utilization patterns over the shared topology.

Contributions.

- We introduce compound-conditioned bioactivity graphs with a shared mechanistic topology and chemical-specific node and edge gates for developmental-toxicity prediction.
- We construct the topology using a six-level AOP-inspired hierarchy and co-activation features based on nPMI and Yule's Q.
- Structured transfer outperforms flat-logit, molecular-only, and random-forest baselines across five held-out splits and all six phenotypes.
- Joint node and edge gating improves AUROC by 0.046 over a static graph, and split-wise visualizations reveal heterogeneous compound-specific assay utilization.

2 Related Work

2.1 Bioactivity descriptors for in vivo toxicity prediction

Using *in vitro* screening data to predict *in vivo* outcomes is well established in predictive toxicology. Combining ToxCast bioactivity with chemical descriptors improves prediction of rat liver histopathology and organ-level toxicity, with hybrid representations generally outperforming either modality alone [Liu et al., 2015, 2017]. This supports the mechanistic premise that structure does not fully capture the biological interactions governing organism-level outcomes [Han et al., 2024].

Experimental assay coverage remains limited, motivating the use of predicted bioactivity. Garcia de Lomana et al. [2021] used assay-specific predictions as descriptors for genotoxic, hepatic, and cardiological endpoints, while the Chemical Checker and its signaturizers infer broader biological signatures for sparsely characterized compounds [Duran-Frigola et al., 2020, Bertoni et al., 2021]. Our Tier-I stage follows this transfer strategy by generating assay profiles for every downstream compound. The distinction lies in their downstream representation: prior approaches primarily consume predicted activities as flat vectors, whereas we organize them as states over a mechanistically structured assay graph.

2.2 Zebrafish developmental toxicity and its computational modeling

Embryonic zebrafish provide organism-level developmental readouts at screening throughput in a vertebrate with substantial human conservation: approximately 70% of human protein-coding genes have a zebrafish orthologue, including 82% of genes with OMIM morbidity annotations [Howe et al., 2013]. This conservation supports transfer from human-target assays to zebrafish phenotypes, although concordance with mammalian teratogenicity is limited by differences in placental exposure, embryonic metabolism, and teleost paralogue subfunctionalization [Ball et al., 2014].

Truong et al. [2014] screened 1,060 ToxCast compounds across 18 developmental endpoints, identifying 487 compounds with significant adverse responses; these data underpin much subsequent modeling. Prior work has emphasized structure-based deep and generative models [Green et al., 2021], multitask learning across zebrafish assays [Lin et al., 2024], and models for specific compound classes [Yogodzinski et al., 2025]. The closest two-tier precedent is Santos et al. [2026], who combine molecular and AOP-anchored pathway graphs through cross-attention to predict endocrine outcomes using 46 receptor assays. We instead retain 457 assays, impose a broader mechanistic topology, and adapt both assay nodes and assay–assay edges for each compound.

2.3 Graph neural networks for molecular property prediction

Message passing is the standard neural approach to molecular property prediction [Gilmer et al., 2017], with widely used variants including GCN, GAT, GIN, and AttentiveFP [Kipf and Welling, 2017, Veličković et al., 2018, Xu et al., 2019, Xiong et al., 2020]. MoleculeNet and directed-MPNN studies established common evaluation practices, including scaffold-based assessment of chemical generalization [Wu et al., 2018, Yang et al., 2019, Bemis and Murcko, 1996]. We evaluate five backbones at both tiers and find that the preferred architecture differs between molecular bioactivity pretraining and conditioned biological-graph learning.

Unlike molecular graphs, our assay graph contains biological measurements as nodes. Its states are predicted by an upstream model and subsequently used to condition node and edge gating, combining pathway-structured message passing with transferred bioactivity supervision.

2.4 Conditioning, multimodal alignment, and learning under partial labels

Feature-wise conditioning modulates computation using auxiliary inputs [Perez et al., 2018]; our gates apply this principle to assay node states and assay–assay edge embeddings. We align molecular, bioactivity, and descriptor views using symmetric InfoNCE [van den Oord et al., 2018, Radford et al., 2021]. Unlike molecular self-supervision [Wang et al., 2022, Hu et al., 2020], this objective aligns heterogeneous representations of the same labeled compound. The masked multi-label regularizer adapts supervised contrastive learning [Khosla et al., 2020] by comparing only jointly observed endpoints, while homoscedastic uncertainty weighting balances tasks of differing difficulty and prevalence [Kendall et al., 2018].

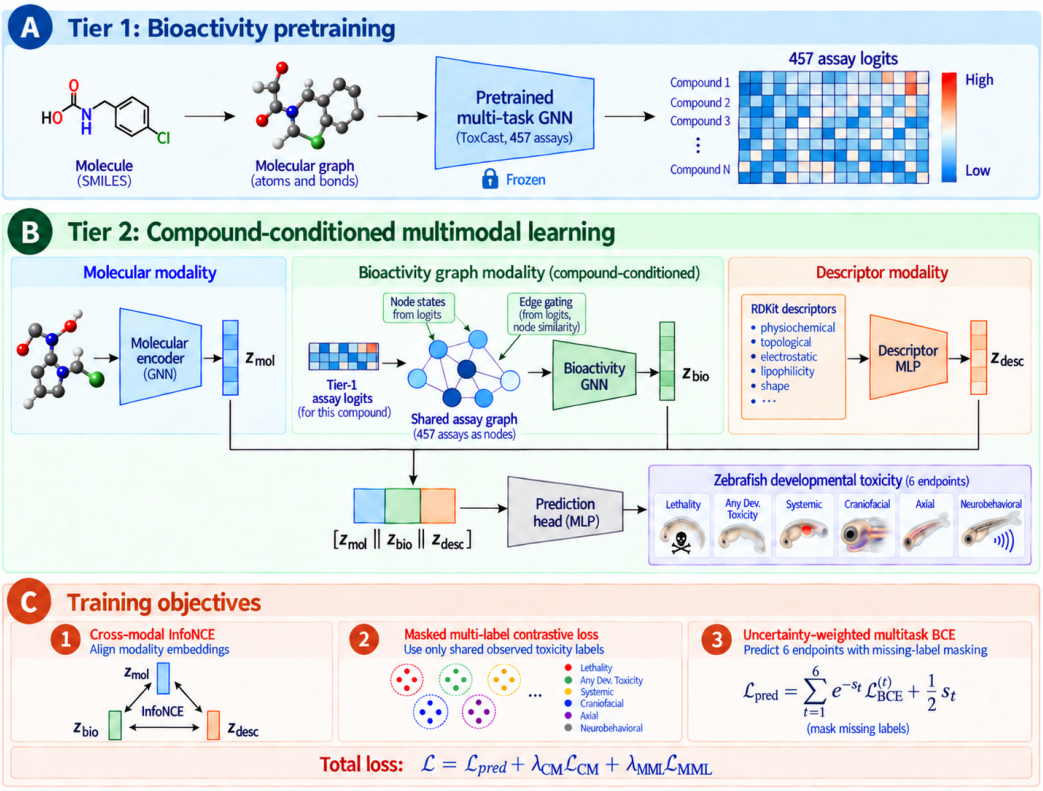


Figure 1: Two-tier framework. Tier-I assay logits condition a shared bioactivity graph through compound-specific node and edge gates; Tier II integrates this graph with molecular and physicochemical representations.

3 Method

3.1 Problem Formulation and Framework Overview

We consider a two-stage problem in which a large, partially observed *in vitro* dataset provides biological supervision for a smaller developmental-toxicity task. Tier I contains molecular graphs G_i, assay activities $\mathbf{a}_i \in \{0,1\}^{457}$, and observation masks $\mathbf{m}_i^A$. Its frozen predictor supplies Tier-II bioactivity logits $\mathbf{b}_i \in \mathbb{R}^{457}$. Tier II additionally contains physicochemical descriptors $\mathbf{d}_i$, endpoint labels $\mathbf{y}_i \in \{0,1\}^K$, and label masks $\mathbf{m}_i^Y$. The objective is

$$f_\Theta\left(G_i, \mathbf{b}_i, \mathbf{d}_i\right) \to \hat{\mathbf{y}}_i, \tag{1}$$

while exploiting molecular, biological, and physicochemical information under sparse and incomplete endpoint supervision.

3.2 Tier-I Multitask Bioactivity Pretraining

Each canonical SMILES is converted to a molecular graph and encoded by a multitask GNN. Its assay logits are

$$\mathbf{b}_i = f_{\text{head}}^I\left(f_\theta^I(G_i)\right) \in \mathbb{R}^{457}. \tag{2}$$

Training uses masked binary cross-entropy so unmeasured compound–assay pairs do not contribute. After model selection, the encoder and head are frozen to generate Tier-II profiles.

3.3 Compound-Conditioned Bioactivity Graph

Rather than consuming Tier-I predictions as a flat vector, we organize them on a shared assay graph $\mathcal{G}_B = (V_B, E_B)$. Each node is a ToxCast assay, and edges encode the mechanistic and co-activation relationships described in Appendix B. For compound c, assay logit $b_j^{(c)}$ initializes node j:

$$\mathbf{h}_j^{(0,c)} = \phi_{\text{in}}\left(b_j^{(c)}\right). \tag{3}$$

Its contribution is adapted by a node gate,

$$\gamma_j^{(c)} = \sigma\Big(g_n\left(b_j^{(c)}\right)\Big), \qquad \tilde{\mathbf{h}}_j^{(0,c)} = \gamma_j^{(c)}\mathbf{h}_j^{(0,c)}. \tag{4}$$

For edge (j,k) with static features $\mathbf{e}_{jk}$, an edge gate conditions its relevance on both connected assay logits:

$$\alpha_{jk}^{(c)} = \sigma\Big(g_e\left[b_j^{(c)} \| b_k^{(c)} \| \mathbf{e}_{jk}\right]\Big), \qquad \tilde{\mathbf{e}}_{jk}^{(c)} = \alpha_{jk}^{(c)}\phi_e\left(\mathbf{e}_{jk}\right). \tag{5}$$

After L message-passing layers over the shared topology, the compound-level bioactivity embedding is

$$\mathbf{z}_c^{\text{bio}} = \text{READOUT}\left(\left\{\mathbf{h}_j^{(L,c)}\right\}_{j=1}^{457}\right). \tag{6}$$

Thus, all compounds share the same biological graph topology, while the importance of assay nodes and assay relationships is dynamically conditioned on the predicted bioactivity profile of each chemical.

3.4 Multimodal Developmental-Toxicity Prediction

Tier II uses a molecular GNN, the conditioned assay graph, and an MLP over approximately 100 RDKit descriptors. Their embeddings are fused by concatenation:

$$\begin{aligned} \mathbf{z}_i^{\text{mol}} &= f_{\text{mol}}(G_i), \quad \mathbf{z}_i^{\text{bio}} = f_{\text{bio}}(\mathcal{G}_B, \mathbf{b}_i), \\ \mathbf{z}_i^{\text{desc}} &= f_{\text{desc}}(\mathbf{d}_i), \quad \hat{\mathbf{y}}_i = f_{\text{pred}}\big([\mathbf{z}_i^{\text{mol}} \| \mathbf{z}_i^{\text{bio}} \| \mathbf{z}_i^{\text{desc}}]\big). \end{aligned} \tag{7}$$

3.5 Representation-Learning Objectives

Modality-specific projection heads map the molecular, bioactivity, and descriptor embeddings into a normalized shared space. A symmetric InfoNCE objective aligns views of the same compound and averages the molecular–bioactivity, molecular–descriptor, and bioactivity–descriptor pairs:

$$\mathcal{L}_{\text{CM}} = \frac{1}{3}\left(\mathcal{L}_{\text{mol,bio}} + \mathcal{L}_{\text{mol,desc}} + \mathcal{L}_{\text{bio,desc}}\right). \tag{8}$$

To exploit incomplete phenotype labels without treating missing entries as negatives, the masked multi-label objective matches latent cosine similarity to phenotype agreement computed only over jointly observed endpoints. It is averaged across the molecular and bioactivity representations. Full definitions of the projection, similarity, agreement, and masking terms are provided in Appendix C.4.

3.6 Multitask Training Objective

To account for differences in task difficulty and uncertainty across developmental-toxicity endpoints, prediction is optimized using a homoscedastic uncertainty-weighted multitask objective:

$$\mathcal{L}_{\text{pred}} = \frac{1}{K}\sum_{k=1}^{K}\left[\exp(-s_k)\mathcal{L}_{\text{WBCE}}^{(k)} + \frac{1}{2}s_k\right], \tag{9}$$

where s_k is a learned task-specific uncertainty parameter and $\mathcal{L}_{\mathrm{WBCE}}^{(k)}$ denotes the masked weighted binary cross-entropy for endpoint k.

The complete optimization objective is

$$\boxed{\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_{\mathrm{CM}}\mathcal{L}_{\mathrm{CM}} + \lambda_{\mathrm{MML}}\mathcal{L}_{\mathrm{MML}}} \tag{10}$$

where λ_{CM} and λ_{MML} control the contribution of the two representation-learning objectives.

4 Experimental Setup

4.1 Datasets and Evaluation Protocol

Tier I contains 9,020 compounds measured across 457 ToxCast assays; unmeasured compound–assay entries are masked. The frozen Tier-I model produces a 457-dimensional logit profile for each of 1,415 Tier-II compounds. Tier-II labels comprise six partially observed zebrafish phenotype groups: lethality, any developmental toxicity, systemic, craniofacial, axial, and neurobehavioral toxicity. Inputs combine the molecular graph, conditioned assay graph, and approximately 100 RDKit descriptors. Appendix A gives full preprocessing and label statistics.

We report five repeated stratified splits with 70/15/15 training, validation, and test proportions. The split seeds are 42, 184, 758, 792, and 7. Validation data are used for early stopping and threshold selection; test compounds remain isolated. The implementation also supports Bemis–Murcko scaffold splits, described in the appendix.

4.2 Model Backbones and Baselines

We evaluate MPNN, GAT, GCN, GIN, and AttentiveFP backbones. Comparisons include a molecular-only multitask GNN, a molecular–bioactivity model without descriptors, the complete multimodal model, and a per-endpoint Random Forest using Morgan fingerprints, descriptors, and Tier-I logits. Architecture sensitivity is reported in Appendix C.1.

4.3 Hyperparameter Optimization and Model Training

Hyperparameters were selected by validation AUROC using Optuna/TPE with median pruning [Akiba et al., 2019]. Adam optimization used component-specific learning rates, 10^{-5} weight decay, plateau-based learning-rate reduction, gradient clipping at 1.0, and early stopping with restoration of the best checkpoint. Endpoint-specific positive weights address imbalance; observation masks exclude missing labels from prediction and representation losses. Appendix C.5 provides the complete search and training protocol.

4.4 Evaluation Metrics

The primary metric is macro-averaged endpoint AUROC. We additionally report AUPRC, balanced accuracy, sensitivity, specificity, F1, G-mean, and Brier score. Endpoint-specific operating thresholds maximize validation G-mean and are then fixed for testing. Results are mean $\pm$ standard deviation across five seeds.

Table 1: Five-seed held-out performance (mean ± standard deviation). Thresholds for balanced accuracy and G-mean are selected on validation data.

Model	AUROC	AUPRC	Bal. Acc.	G-mean
Molecular-only GAT	0.718 ± 0.040	0.317 ± 0.065	0.653 ± 0.040	0.645 ± 0.044
Molecular + Bioactivity GAT	0.751 ± 0.024	0.321 ± 0.042	0.686 ± 0.025	0.677 ± 0.027
Random Forest	0.730 ± 0.016	0.310 ± 0.046	0.665 ± 0.025	0.655 ± 0.029
Full multimodal GAT	$\mathbf{0.786 \pm 0.020}$	$\mathbf{0.342 \pm 0.041}$	$\mathbf{0.717 \pm 0.037}$	$\mathbf{0.706 \pm 0.043}$

Table 2: Comparison of flat and graph-structured Tier-I bioactivity transfer across five matched held-out seeds. Values are mean ± standard deviation.

Bioactivity representation	AUROC	AUPRC	Bal. Acc.	G-mean
Flat Tier-I logits + MLP	0.723 ± 0.034	0.304 ± 0.057	0.639 ± 0.035	0.629 ± 0.037
Structured bioactivity graph	$\mathbf{0.786 \pm 0.020}$	$\mathbf{0.342 \pm 0.041}$	$\mathbf{0.717 \pm 0.037}$	$\mathbf{0.706 \pm 0.043}$

5 Results

5.1 Bioactivity-Informed Multimodal Learning Improves Developmental Toxicity Prediction

We first evaluated whether incorporating intermediate bioactivity information improves developmental toxicity prediction beyond molecular structure alone. Table 1 summarizes performance across five independent held-out train–validation–test splits for the GAT-based models and the Random Forest (RF) baseline. Results are reported as mean ± standard deviation across the five evaluation seeds.

The molecular-only GAT achieved a mean AUROC of 0.718 ± 0.040. Incorporating the Tier-I bioactivity representation increased AUROC to 0.751 ± 0.024, corresponding to an absolute improvement of 0.033. The full multimodal model achieved the highest mean AUROC, 0.786 ± 0.020, representing improvements of 0.068 over the molecular-only model and 0.035 over the molecular–bioactivity model. The full model also outperformed the RF baseline (0.730 ± 0.016), yielding an absolute AUROC improvement of 0.056.

A similar pattern was observed for threshold-dependent metrics. Following threshold selection on the validation set, the full multimodal model achieved a balanced accuracy of 0.717 ± 0.037 and a G-mean of 0.706 ± 0.043, compared with 0.653 ± 0.040 and 0.645 ± 0.044, respectively, for the molecular-only model. These results indicate that the improvement is not restricted to overall ranking performance, but is also reflected in a better balance between positive- and negative-class prediction under class imbalance.

5.2 Structured Bioactivity Transfer Outperforms a Flat Bioactivity Representation

To determine whether the improvement from Tier-I bioactivity arose simply from providing predicted assay responses or from explicitly modeling their relationships, we compared the full bioactivity-graph model with a flat bioactivity baseline. In the flat condition, the Tier-I assay logits were processed by an MLP rather than propagated through the mechanistically structured assay graph, while the remaining Tier-II architecture, training objectives, and evaluation protocol were retained.

Table 2 shows that, across five matched held-out seeds, the flat bioactivity model achieved an AUROC of 0.723 ± 0.034, compared with 0.786 ± 0.020 for the full graph-based model. The graph representation also improved AUPRC from 0.304 ± 0.057 to 0.342 ± 0.041, balanced accuracy from 0.639 ± 0.035 to 0.717 ± 0.037, and G-mean from 0.629 ± 0.037 to 0.706 ± 0.043.

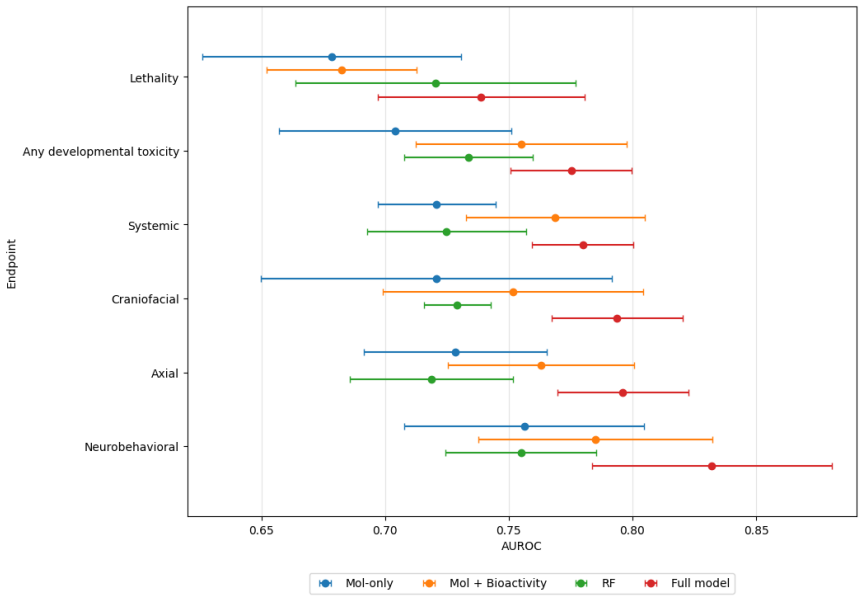


Figure 2: Endpoint-level AUROC across five held-out evaluations. Points show means and error bars show one standard deviation.

These results indicate that the benefit of Tier-I transfer is not explained solely by access to predicted bioactivity features. Explicitly organizing the assay profile as a relational biological graph provides additional predictive value beyond a flat vector representation, supporting the use of structured intermediate supervision for downstream developmental-toxicity prediction.

The performance advantage of the structured representation motivated further inspection of the learned compound-specific bioactivity graphs. Representative training, validation, and held-out test compounds showed distinct patterns of assay and interaction utilization despite sharing the same underlying assay topology (Appendix D.2; Figure 6). These visualizations provide a qualitative view of how the adaptive graph encoder reweights intermediate bioactivity information on a per-compound basis.

5.3 Performance Improves Across Individual Developmental-Toxicity Phenotypes

We next examined whether the gains of the proposed framework were consistent across individual developmental-toxicity phenotypes rather than being driven by a small subset of endpoints. Figure 2 summarizes endpoint-level AUROC across five independent held-out evaluations for the molecular-only GAT, molecular–bioactivity GAT, Random Forest (RF), and full multimodal GAT models. Points indicate the mean AUROC across the five seeds, with horizontal error bars denoting the corresponding standard deviation.

The full multimodal GAT achieved the highest mean AUROC for all six phenotype groups. The strongest performance was observed for neurobehavioral toxicity (0.832 ± 0.048), followed by axial toxicity (0.796 ± 0.027) and craniofacial toxicity (0.794 ± 0.026). Lethality remained the most challenging endpoint, although the full model improved performance to 0.739 ± 0.042.

A progressive improvement was also observed when moving from molecular structure alone to bioactivity-informed and multimodal representations. Relative to the molecular-only GAT, incorporation of the Tier-I bioactivity representation increased AUROC for any developmental toxicity (0.704 ± 0.047 to 0.755 ± 0.043), systemic toxicity (0.721 ± 0.024 to 0.769 ± 0.036), axial toxicity

Table 3: Gating ablation across five matched held-out seeds (mean $\pm$ standard deviation).

Configuration	Node	Edge	AUROC	AUPRC	Bal. Acc.	G-mean
Static bioactivity graph	–	–	0.740 ± 0.025	0.324 ± 0.033	0.659 ± 0.037	0.653 ± 0.038
Node gating only	✓	–	0.729 ± 0.057	0.270 ± 0.072	0.667 ± 0.042	0.654 ± 0.039
Edge gating only	–	✓	0.722 ± 0.036	0.270 ± 0.063	0.651 ± 0.033	0.636 ± 0.025
Full model (reference)	✓	✓	$\mathbf{0.786 \pm 0.020}$	$\mathbf{0.342 \pm 0.041}$	$\mathbf{0.717 \pm 0.037}$	$\mathbf{0.706 \pm 0.043}$

(0.728 ± 0.037 to 0.763 ± 0.038), and neurobehavioral toxicity (0.756 ± 0.048 to 0.785 ± 0.047).

The full multimodal formulation provided additional gains across all phenotype groups, with particularly notable improvements over the molecular–bioactivity model for lethality ($+0.056$ AUROC), craniofacial toxicity ($+0.042$), and neurobehavioral toxicity ($+0.047$). The RF baseline was generally competitive with the molecular-only model but was consistently outperformed by the full multimodal GAT. Together, these results indicate that the overall macro-level improvement reflects broad gains across developmental phenotypes rather than dominance by a single endpoint.

5.4 Joint Node and Edge Gating Achieves the Best Performance

Table 3 summarizes the five-seed held-out performance of the gating ablations. The previously evaluated full model, which jointly applies compound-specific node and edge gating, achieved the highest mean AUROC (0.786 ± 0.020). Removing both adaptive gates and retaining a static bioactivity graph reduced performance to 0.740 ± 0.025, corresponding to an absolute decrease of 0.046 AUROC.

Neither gating mechanism alone recovered the performance of the fully gated model: node-only and edge-only variants achieved 0.729 ± 0.057 and 0.722 ± 0.036 AUROC, respectively, giving full-model improvements of 0.057 and 0.064. Threshold-dependent metrics followed the same pattern. The full model achieved 0.717 ± 0.037 balanced accuracy and 0.706 ± 0.043 G-mean, compared with 0.659 ± 0.037 and 0.653 ± 0.038 for the static graph; both single-gate variants also remained below the full model.

Because neither gate individually exceeded the static graph in mean AUROC, the adaptive mechanisms appear complementary rather than independently additive: node gating modulates assay responses, while edge gating modulates compound-specific assay–assay interactions. Joint gating therefore adapts both assay states and their relational structure to each compound.

6 Conclusion

We introduced compound-conditioned bioactivity graphs for developmental-toxicity prediction, replacing flat assay vectors with chemically adaptive states over a shared mechanistic topology. The full model achieved 0.786 ± 0.020 AUROC across five held-out splits, outperforming molecular structure alone (0.718 ± 0.040), molecular–bioactivity fusion (0.751 ± 0.024), and a random forest using the same modalities (0.730 ± 0.016). Gains extended across all six developmental phenotypes.

Joint node and edge gating improved AUROC by 0.046 over a static assay graph, whereas neither gate alone exceeded it. Thus, transfer benefits from adapting both assay states and their relationships, connecting high-throughput screening with AOP-inspired phenotype prediction [Richard et al., 2016, Ankley et al., 2010] and complementing recent two-tier graph models [Santos et al., 2026]. Limitations include non-causal gates and topology, Tier-I error, and cross-species transfer [Truong et al., 2014]; future work should pursue external validation and dose-aware modeling.

AI Use Statement

Generative AI tools were used to assist with manuscript organization, drafting and language editing, LaTeX formatting, literature searches, bibliography verification, consistency checks, and the wording of qualitative interpretations. They were not used to generate experimental data, preprocess datasets, implement or train the reported models, calculate evaluation metrics, or autonomously determine the scientific conclusions. The reported results were supplied by the authors, and the authors reviewed and verified all AI-assisted text, citations, calculations, and claims. The authors take responsibility for the complete content of this submission.

Reproducibility Statement

Section 3 specifies the two-tier model, adaptive gates, representation objectives, and multitask loss. Section 4 reports dataset composition, split construction, training procedures, model selection, metrics, and all five held-out seeds. The appendix details chemical preprocessing (Appendix A.4), assay-graph construction (Appendix B), backbone comparisons (Appendix C.1), representation objectives (Appendix C.4), and optimization and checkpoint selection (Appendix C.5). These materials document the transformations, model components, objectives, evaluation protocol, and random seeds required to reproduce the reported experiments.

Ethics Statement

This work follows the ICLR Code of Ethics and uses previously collected chemical bioactivity and zebrafish developmental-toxicity data; no new experiments involving human participants or animals were conducted. The proposed model is intended to support research and chemical-prioritization workflows rather than replace experimental toxicology or serve as a standalone regulatory decision system. False-negative and false-positive predictions may have safety and resource-allocation consequences, and performance may be affected by assay coverage, endpoint imbalance, chemical-domain shift, and cross-species differences. Compound-specific gates and graph visualizations reflect model utilization of assay information and must not be interpreted as causal pathway evidence. Responsible use therefore requires external validation, uncertainty-aware interpretation, and confirmation through appropriate experimental studies.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

Gerald T. Ankley, Richard S. Bennett, Russell J. Erickson, Dale J. Hoff, Michael W. Hornung, Rodney D. Johnson, David R. Mount, John W. Nichols, Christine L. Russom, Patricia K. Schmieder, Jose A. Serrrano, Joseph E. Tietge, and Daniel L. Villeneuve. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3):730–741, 2010.

Jonathan S. Ball, Donald B. Stedman, Jedd M. Hillegass, Cindy X. Zhang, Julie Panzica-Kelly, Aleasha Coburn, Brian P. Enright, Belen Tornesi, Hamid R. Amouzadeh, Malcolm Hetheridge, Anne-Lee Gustafson, and Karen A. Augustine-Rauch. Fishing for teratogens: A consortium effort for a harmonized zebrafish developmental toxicology assay. *Toxicological Sciences*, 139(1): 210–219, 2014. doi: 10.1093/toxsci/kfu017.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

Martino Bertoni, Miquel Duran-Frigola, Pau Badia-i Mompel, Eduardo Pauls, Modesto Orozco-Ruiz, Oriol Guitart-Pla, Víctor Alcalde, Víctor M. Diaz, Antonio Berenguer-Llergo, Isabelle Brun-Heath, Núria Villegas, Antonio García de Herreros, and Patrick Aloy. Bioactivity descriptors for uncharacterized chemical compounds. *Nature Communications*, 12(1):3932, 2021.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen, 2009. Gunter Narr Verlag.

Eugene G. DeMaster, Frances N. Shirota, and Herbert T. Nagasawa. The metabolic activation of cyanamide to an inhibitor of aldehyde dehydrogenase is catalyzed by catalase. *Biochemical and Biophysical Research Communications*, 122(1):358–365, 1984. doi: 10.1016/0006-291X(84) 90483-2.

Eugene G. DeMaster, Barbara Redfern, and Herbert T. Nagasawa. Mechanisms of inhibition of aldehyde dehydrogenase by nitroxyl, the active metabolite of the alcohol deterrent agent cyanamide. *Biochemical Pharmacology*, 55(12):2007–2015, 1998. doi: 10.1016/S0006-2952(98)00080-X.

Miquel Duran-Frigola, Eduardo Pauls, Oriol Guitart-Pla, Martino Bertoni, Víctor Alcalde, David Amat, Teresa Juan-Blanco, and Patrick Aloy. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nature Biotechnology*, 38(9):1087–1096, 2020.

Marina Garcia de Lomana, Andrea Morger, Ulf Norinder, Roland Buesen, Robert Landsiedel, Andrea Volkamer, Johannes Kirchmair, and Miriam Mathea. ChemBioSim: Enhancing conformal prediction of in vivo toxicity by use of predicted bioactivities. *Journal of Chemical Information and Modeling*, 61(7):3255–3272, 2021.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, 2017.

Adrian J. Green, Martin J. Mohlenkamp, Jhuma Das, Meenal Chaudhari, Lisa Truong, Robyn L. Tanguay, and David M. Reif. Leveraging high-throughput screening data, deep neural networks, and conditional generative adversarial networks to advance predictive toxicology. *PLOS Computational Biology*, 17(7):e1009135, 2021.

Daniel R. Hallinger, Hayley B. Lindsay, Katie Paul Friedman, Danielle A. Suarez, and Steven O. Simmons. Respirometric screening and characterization of mitochondrial toxicants within the ToxCast phase I and II chemical libraries. *Toxicological Sciences*, 176(1):175–192, 2020. doi: 10.1093/toxsci/kfaa059.

Peiling Han, Xuehua Li, Jingyuan Yang, Yuxuan Zhang, and Jingwen Chen. Advancing toxicity predictions: A review on in vitro to in vivo extrapolation in next-generation risk assessment. *Environment & Health*, 2(7):499–513, 2024. doi: 10.1021/envhealth.4c00043.

Kerstin Howe et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, 496(7446):498–503, 2013. doi: 10.1038/nature12111.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Run-Hsin Lin, Pinpin Lin, Chia-Chi Wang, and Chun-Wei Tung. A novel multitask learning algorithm for tasks with distinct chemical space: Zebrafish toxicity prediction as an example. *Journal of Cheminformatics*, 16:91, 2024. doi: 10.1186/s13321-024-00891-4.

Jie Liu, Kamel Mansouri, Richard S. Judson, Matthew T. Martin, Huixiao Hong, Minjun Chen, Xiaowei Xu, Russell S. Thomas, and Imran Shah. Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. *Chemical Research in Toxicology*, 28(4):738–751, 2015.

Jie Liu, Grace Patlewicz, Antony J. Williams, Russell S. Thomas, and Imran Shah. Predicting organ toxicity using in vitro bioactivity data and chemical structure. *Chemical Research in Toxicology*, 30(11):2046–2059, 2017.

Peter P. Major, Eileen M. Egan, David J. Herrick, and Donald W. Kufe. Effect of ARA-C incorporation on deoxyribonucleic acid synthesis in cells. *Biochemical Pharmacology*, 31(18):2937–2940, 1982. doi: 10.1016/0006-2952(82)90266-0.

Tsuyoshi Nakanishi, Jun-ichi Nishikawa, Youhei Hiromori, et al. Trialkyltin compounds bind retinoid X receptor to alter human placental endocrine functions. *Molecular Endocrinology*, 19(10): 2502–2516, 2005. doi: 10.1210/me.2004-0397.

Eva Oliveira, Marta Casado, Demetrio Raldúa, Amadeu Soares, Carlos Barata, and Benjamin Piña. Retinoic acid receptors' expression and function during zebrafish early development. *Journal of Steroid Biochemistry and Molecular Biology*, 138:143–151, 2013. doi: 10.1016/j.jsbmb.2013.03.011.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Ann M. Richard, Richard S. Judson, Keith A. Houck, Christopher M. Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T. Martin, John F. Wambaugh, Thomas B. Knudsen, Jayaram Kancherla, Kamel Mansouri, Grace Patlewicz, Antony J. Williams, Stephen B. Little, Kevin M. Crofton, and Russell S. Thomas. ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chemical Research in Toxicology*, 29(8):1225–1251, 2016.

Sebastián Sánchez-Fortún, María Teresa Llorente, and Argelia Castaño. Genotoxic effects of selected biocides on RTG-2 fish cells by means of a modified fast micromethod assay. *Aquatic Toxicology*, 73(1):55–64, 2005. doi: 10.1016/j.aquatox.2005.03.002.

Eder Soares de Almeida Santos, Gustavo Felizardo Santos Sandes, Artur Christian Garcia da Silva, Holli-Joi Martin, Eugene N. Muratov, Rodolpho de Campos Braga, and Bruno Junior Neves. Multimodal cross-attentive graph-based framework for predicting in vivo endocrine disruptors. *Advanced Science*, 13(21):e19897, 2026. doi: 10.1002/advs.202519897.

Raymond R. Tice, Christopher P. Austin, Robert J. Kavlock, and John R. Bucher. Improving the human hazard characterization of chemicals: A Tox21 update. *Environmental Health Perspectives*, 121(7):756–765, 2013.

Lisa Truong, David M. Reif, Lindsey St Mary, Mitra C. Geier, Hao D. Truong, and Robert L. Tanguay. Multidimensional in vivo hazard assessment using zebrafish. *Toxicological Sciences*, 137(1): 212–233, 2014. doi: 10.1093/toxsci/kft235.

U.S. Environmental Protection Agency. Toxicological review of trimethylbenzenes. Technical Report EPA/635/R-16/161Fa, Integrated Risk Information System, U.S. Environmental Protection Agency, 2016.

Ricardo Valcárcel, Heinz Holz, Carmen G. Jiménez, Diego Barettino, and Hendrik G. Stunnenberg. Retinoid-dependent in vitro transcription mediated by the RXR/RAR heterodimer. *Genes & Development*, 8(24):3068–3079, 1994. doi: 10.1101/gad.8.24.3068.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Christopher Yogodzinski, Joshua S. Harris, Thomas R. Lane, Morgan Barnes, Patricia A. Vignaux, Renuka Raman, Lisa Truong, Robyn L. Tanguay, Seth Kullman, and Sean Ekins. Machine learning modeling of zebrafish toxicity endpoints after exposure to PROTACs. *Toxicological Sciences*, 2025. doi: 10.1093/toxsci/kfaf162.

G. Udny Yule. On the association of attributes in statistics: With illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A*, 194: 257–319, 1900. doi: 10.1098/rsta.1900.0019.

Qun-Fang Zhang, Ying-Wen Li, Zhi-Hao Liu, Qi-Liang Chen, et al. Exposure to mercuric chloride induces developmental damage, oxidative stress and immunotoxicity in zebrafish embryos-larvae. *Aquatic Toxicology*, 181:76–85, 2016. doi: 10.1016/j.aquatox.2016.10.029.

A Datasets and Preprocessing

A.1 Tier-I ToxCast Bioactivity Dataset

Chemical bioactivity measurements were obtained from the U.S. Environmental Protection Agency (EPA) Toxicity Forecaster (ToxCast) program through InvitroDB v4.3. The complete database contained measurements spanning 1,585 *in vitro* assays. To obtain a sufficiently supported multitask learning dataset, assays were retained only when activity measurements were available for at least 1,000 compounds. Background, control, and non-mechanistic measurement assays were removed. Zebrafish developmental-toxicity assays were also excluded to prevent direct overlap between the Tier-I pretraining targets and the downstream Tier-II developmental-toxicity outcomes.

Following assay filtering, the retained assay panel comprised 457 assays distributed across 32 biological target families (Figure 3). The assay panel covered a broad range of biological processes, including nuclear-receptor signaling, cell-cycle regulation, DNA binding, cytokine signaling, metabolic enzymes, ion channels, and other mechanistically relevant target classes.

Assay-level activity measurements were retrieved using assay endpoint identifiers (AEIDs) and merged into a unified compound-by-assay activity matrix. Hit-call annotations were represented as binary labels, with $y_{ij} = 1$ indicating an active response of compound i in assay j and $y_{ij} = 0$ indicating an experimentally observed inactive response. Unobserved compound–assay combinations were retained as missing values rather than being interpreted as inactive measurements.

Chemical structures were mapped to SMILES representations using DSSTox substance identifiers (DTXSIDs). Compounds lacking a valid or RDKit-parsable structure were excluded. Following structure standardization and quality control, the Tier-I dataset contained 9,020 unique compounds.

The resulting bioactivity matrix was highly incomplete and imbalanced. Approximately 56.4% of compound–assay entries were unobserved, reflecting the heterogeneous testing coverage of the ToxCast program. Among experimentally observed entries, only 11.5% were active responses. Consequently, Tier-I learning required simultaneous handling of substantial label sparsity and positive-class imbalance.

A.2 Tier-II Zebrafish Developmental-Toxicity Dataset

The downstream developmental-toxicity dataset was constructed by integrating zebrafish screening data from Truong et al. (2014) and the harmonized resource reported by Yogodzinski et al. (2025). Following chemical standardization, canonical structure matching, and removal of duplicate compounds, the final Tier-II dataset contained 1,415 unique chemicals.

Of these compounds, 1,042 (73.6%) originated from the Truong et al. dataset and 373 (26.4%) were contributed by the additional curated studies represented in the Yogodzinski et al. resource.

Developmental phenotypes were organized into six biologically motivated prediction targets: lethality, any developmental toxicity, systemic toxicity, craniofacial toxicity, axial toxicity, and neurobehavioral toxicity. Because not every compound was evaluated for every phenotype, endpoint labels were represented together with an observation mask. Missing endpoint labels were excluded from both predictive and label-aware representation-learning objectives rather than being assigned to the negative class.

A.3 Label Sparsity and Class Imbalance

The two learning stages exhibit distinct forms of sparse supervision. In Tier I, 56.4% of compound–assay combinations were unobserved, and only 11.5% of experimentally observed responses were

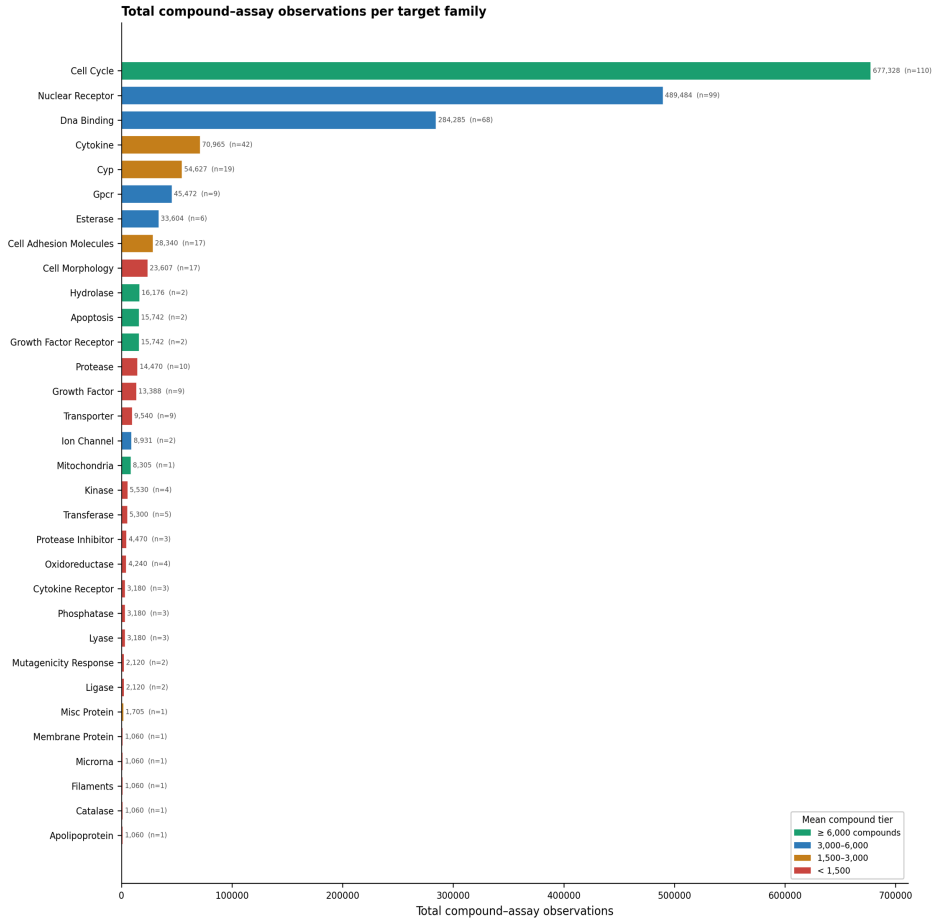


Figure 3: **Coverage of the Tier-I ToxCast assay panel across biological target families.** Bars represent the total number of experimentally observed compound–assay measurements within each target family after assay filtering. Numbers beside each bar indicate the number of retained assays belonging to that family. Colors denote the mean compound coverage of assays within each family. The retained assay panel spans a heterogeneous range of biological processes while exhibiting substantial differences in experimental coverage across target classes.

active. The model therefore learns from a highly sparse and predominantly negative multitask label matrix.

Tier II contains substantially fewer compounds and heterogeneous phenotype coverage. Positive prevalence also differs across developmental-toxicity endpoints. Missing phenotype labels were explicitly masked throughout training and evaluation. Consequently, performance was evaluated using threshold-independent metrics such as AUROC and AUPRC together with class-balanced threshold-dependent measures including balanced accuracy and G-mean.

A.4 Chemical Preprocessing

Chemical structures were represented using SMILES strings and processed with RDKit. Compounds with invalid or unparsable SMILES were excluded. Valid molecules were converted into molecular graphs in which atoms were represented as nodes and chemical bonds as edges. Atom features encoded element type, degree, formal charge, hybridization state, and aromaticity, while bond

features encoded bond type, conjugation, ring membership, and stereochemistry. Each bond was represented bidirectionally to support message passing in the graph neural network.

In parallel, a set of 100 two-dimensional RDKit physicochemical descriptors was calculated for each valid compound and used as a separate descriptor modality in the Tier-II model. Descriptor calculation failures and non-finite values were replaced with zero, and descriptor values were standardized prior to model training.

B Bioactivity Graph Construction

B.1 Overview

To provide Tier II with a biologically structured representation of intermediate chemical activity, we constructed a mechanistically constrained assay graph from the retained ToxCast assay panel. Each node corresponds to an individual ToxCast assay, while edges encode plausible relationships between assays based on biological annotation and empirical co-activation patterns.

The graph contains two complementary sources of information. First, assay annotations define a shared mechanistic topology that is common to all compounds. Second, Tier-I assay predictions provide compound-specific biological states that are propagated over this topology during Tier-II learning. Consequently, the framework does not construct an independent graph topology for each chemical. Instead, compounds share the same assay graph while their node states and learned node- and edge-gating weights vary according to the predicted bioactivity profile of each compound.

B.2 AOP-Inspired Assay Nodes and Mechanistic Hierarchy

Each retained ToxCast assay was represented as one node in the bioactivity graph. Rather than treating assays as an unordered collection of experimental measurements, we organized them into a coarse mechanistic hierarchy designed to mirror the directional logic of an adverse outcome pathway (AOP), in which an upstream molecular interaction propagates through intermediate biological events toward increasingly downstream cellular responses.

Assay annotations were derived from EPA invitroDB metadata, including assay function type, intended target type, biological-process target, intended target family, signal direction, and whether the assay measured cell viability. These annotations were used to assign each assay to one of six mechanistic levels spanning molecular interaction to downstream phenotype:

- **L0 – Binding or molecular interaction:** assays measuring direct interactions with molecular targets, such as receptor binding, ligand binding, biochemical activity, or enzyme-level interactions. These assays represent the most upstream events in the hierarchy and are analogous to molecular-initiating-event-like activity.
- **L1 – Transcriptional or reporter response:** assays measuring receptor activation, agonism/antagonism, reporter-gene activity, or related transcriptional responses occurring downstream of an initial molecular interaction.
- **L2 – Gene-expression response:** assays reflecting gene-expression changes that propagate the initial molecular signal into broader cellular regulatory responses.
- **L3 – Protein or signaling response:** assays capturing kinase activity, phosphorylation, cytokine signaling, protein-level regulation, and related intracellular signaling events.
- **L4 – Cellular phenotype:** assays measuring higher-order cellular responses such as cell-cycle disruption, proliferation, migration, morphology, or other phenotypic alterations.

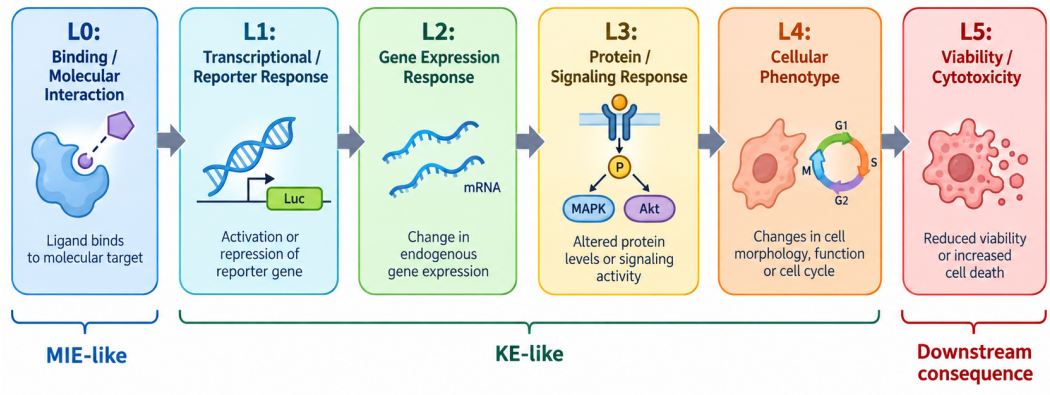


Figure 4: **AOP-inspired mechanistic hierarchy used to organize ToxCast assay nodes in the bioactivity graph.** Retained assays were assigned to six coarse biological levels spanning upstream molecular interactions to downstream cellular consequences: L0, binding or molecular interaction; L1, transcriptional or reporter response; L2, gene-expression response; L3, protein or signaling response; L4, cellular phenotype; and L5, viability or cytotoxicity. The hierarchy is designed to approximate the directional organization of an adverse outcome pathway, with L0 representing molecular-initiating-event (MIE)-like activity, L1–L4 representing progressively downstream key-event (KE)-like responses, and L5 representing a downstream cellular consequence. This hierarchy is used to constrain candidate assay–assay connections in the mechanistic bioactivity graph and should be interpreted as an AOP-inspired computational organization rather than a direct mapping of every assay to a formally curated AOP.

- **L5 – Viability or cytotoxicity:** assays measuring cell viability, apoptosis, necrosis, or generalized cytotoxicity, representing the most downstream cellular consequence in the hierarchy.

The hierarchy should therefore be interpreted as an *AOP-inspired computational organization* of the assay space rather than as a one-to-one mapping of every ToxCast assay to a formally curated molecular initiating event (MIE), key event (KE), or adverse outcome (AO). In this representation, L0 captures MIE-like molecular interactions, L1–L3 represent progressively downstream KE-like molecular and signaling responses, and L4–L5 represent higher-order cellular consequences. This organization provides the graph with an explicit notion of biological directionality while retaining the broader coverage of the ToxCast assay panel.

Biological family information was retained in parallel with hierarchy level. In particular, nuclear-receptor assays were divided into receptor-specific subfamilies, including estrogen receptor (ER), androgen receptor (AR), peroxisome proliferator-activated receptor (PPAR), pregnane X receptor (PXR), farnesoid X receptor (FXR), retinoic acid receptor (RAR), retinoid X receptor (RXR), and related receptor groups. Non-nuclear-receptor assays retained their broader target-family annotation. Consequently, each assay node is characterized by both *what biological system it measures* and *where that measurement lies along the upstream-to-downstream mechanistic hierarchy*.

B.3 Mechanistically Constrained Edge Construction

Candidate assay–assay edges were defined using both biological family and hierarchy information rather than assuming a fully connected graph. For assays i and j, a candidate connection was retained when the assays belonged to the same biological family and occupied either the same or adjacent mechanistic hierarchy levels. This allows local transitions between related biological events, for example from receptor activation to transcriptional response or from transcriptional activity to downstream signaling events.

Additional connections were permitted between high-level phenotypic assays and viability or cytotoxicity assays in order to represent broad cytotoxic effects that may influence multiple downstream biological processes. The resulting candidate topology therefore imposes a mechanistic prior while avoiding unrestricted connectivity among all assay pairs.

B.4 Assay Co-activation Edge Features

For each candidate edge (i, j), empirical assay co-activation statistics were computed from binarized Tier-I assay predictions. A 2×2 contingency table was constructed from the joint firing states of the two assays,

$$\begin{array}{c|cc} & A_j = 1 & A_j = 0 \\ \hline A_i = 1 & n_{11} & n_{10} \\ A_i = 0 & n_{01} & n_{00} \end{array},$$

where n_{11} denotes the number of compounds for which both assays were active, n_{10} and n_{01} denote discordant activation, and n_{00} denotes joint inactivity.

Two association measures were retained as edge attributes: normalized pointwise mutual information (nPMI) [Bouma, 2009] and Yule's Q [Yule, 1900]. For assay pair (i, j),

$$\mathrm{PMI}_{ij} = \log\left(\frac{P(A_i = 1, A_j = 1)}{P(A_i = 1)P(A_j = 1)}\right),$$

and

$$\mathrm{nPMI}_{ij} = \frac{\mathrm{PMI}_{ij}}{-\log P(A_i = 1, A_j = 1)}.$$

Yule's Q was computed directly from the 2×2 contingency table as

$$Q_{ij} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}}.$$

The resulting static edge-feature vector was

$$\mathbf{e}_{ij} = [\mathrm{nPMI}_{ij}, Q_{ij}].$$

These quantities encode complementary information: nPMI measures enrichment of joint assay activation relative to independence, whereas Yule's Q captures the direction and strength of binary association.

In the implemented pipeline, assay reliability was derived from validation AUROC, G-mean, and the separation of probability distributions associated with correct and incorrect predictions. The resulting quantities were combined and bounded to the interval $[0, 1]$, such that poorly discriminating or poorly separated assays exerted less influence on the correction procedure.

B.5 Compound-Conditioned Graph Representation

The constructed assay topology and edge-association attributes are shared across compounds, but each compound induces a distinct biological state on this graph.

For compound c, the frozen Tier-I model produces an assay bioactivity vector

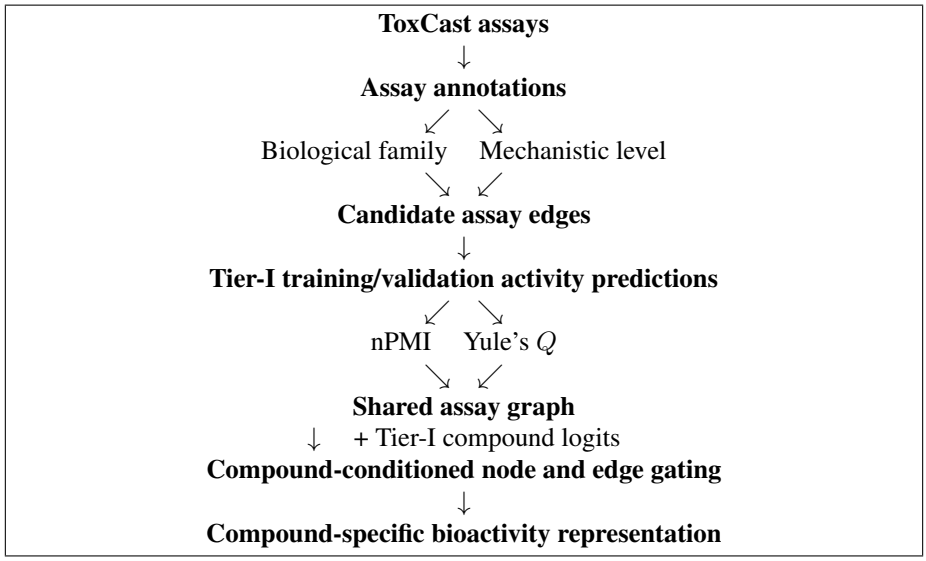


Figure 5: Overview of assay-graph construction and compound-specific instantiation. Assay annotations constrain candidate connectivity, training and validation predictions define association features, and Tier-I compound logits condition node and edge gates on the resulting shared assay graph.

$$\mathbf{b}^{(c)} = \left[b_1^{(c)}, \ldots, b_M^{(c)}\right],$$

where M is the number of retained assays and $b_j^{(c)}$ denotes the Tier-I predicted activity for assay j. These compound-specific assay activities initialize the node states used by the Tier-II bioactivity encoder.

A learned node gate determines the contribution of each assay,

$$\gamma_j^{(c)} = \sigma\left(g_{\text{node}}\left(b_j^{(c)}\right)\right),$$

giving

$$\widetilde{\mathbf{h}}_j^{(c)} = \gamma_j^{(c)} \mathbf{h}_j^{(c)}.$$

Similarly, each candidate assay edge receives a compound-dependent gate,

$$\alpha_{ij}^{(c)} = \sigma\left(g_{\text{edge}}\left(b_i^{(c)} \| b_j^{(c)} \| \text{nPMI}_{ij} \| Q_{ij}\right)\right).$$

The corresponding edge representation becomes

$$\widetilde{\mathbf{e}}_{ij}^{(c)} = \alpha_{ij}^{(c)} \phi_e\left(\mathbf{e}_{ij}\right).$$

The model therefore combines a *shared mechanistically constrained assay graph* with *compound-specific biological states and learned gating*. Two compounds are represented over the same underlying assay network but may emphasize different assay nodes and assay–assay relationships depending on their Tier-I bioactivity profiles.

Overall, graph construction proceeds in three stages: (i) biological annotations define a constrained assay topology; (ii) Tier-I co-activation patterns provide static assay–assay association features; and

Table 4: Tier-I held-out test performance across five graph neural network backbones. Values are mean ± standard deviation across five random seeds. Threshold-dependent metrics correspond to the default decision threshold. The exported evaluation files contained 457 evaluable assay tasks per run.

Backbone	AUROC	Bal. Acc.	Sensitivity	Specificity	F1	G-mean
MPNN	0.798 ± 0.004	0.720 ± 0.005	0.648 ± 0.019	$\mathbf{0.792 \pm 0.015}$	0.405 ± 0.006	0.710 ± 0.006
GAT	0.811 ± 0.004	0.739 ± 0.005	0.799 ± 0.015	0.679 ± 0.015	0.390 ± 0.006	0.731 ± 0.005
GCN	0.812 ± 0.003	0.737 ± 0.004	$\mathbf{0.827 \pm 0.019}$	0.648 ± 0.020	0.382 ± 0.006	0.726 ± 0.005
GIN	$\mathbf{0.821 \pm 0.002}$	$\mathbf{0.751 \pm 0.002}$	0.802 ± 0.016	0.700 ± 0.016	0.404 ± 0.006	$\mathbf{0.744 \pm 0.002}$
AttentiveFP	0.812 ± 0.002	0.741 ± 0.003	0.802 ± 0.032	0.680 ± 0.031	0.392 ± 0.007	0.733 ± 0.004

(iii) Tier-II node and edge gates adapt this shared topology to the predicted bioactivity profile of each compound.

C Detailed Model Architecture and Training

C.1 Graph Neural Network Backbone Evaluation

Rather than fixing a single graph architecture *a priori*, we evaluated five representative graph neural network backbones at both stages of the framework: Message Passing Neural Networks (MPNN), Graph Attention Networks (GAT), Graph Convolutional Networks (GCN), Graph Isomorphism Networks (GIN), and AttentiveFP. This comparison was performed separately for Tier I and Tier II because the two stages operate on fundamentally different graph structures. Tier I learns directly from molecular graphs, whereas Tier II combines a molecular graph with a structured assay–bioactivity graph whose node states and edge importance vary across compounds.

The backbone comparison therefore served as an architecture-sensitivity analysis rather than assuming that the same message-passing mechanism would be optimal for both molecular representation learning and compound-conditioned bioactivity propagation.

C.1.1 Tier-I Backbone Comparison

Across the Tier-I molecular bioactivity task, all five graph architectures achieved mean test AUROC values close to or above 0.80 (Table 4). GIN produced the strongest overall performance, achieving an AUROC of 0.821 ± 0.002, balanced accuracy of 0.751 ± 0.002, and G-mean of 0.744 ± 0.002 across five seeds. GCN achieved the highest sensitivity (0.827 ± 0.019), whereas MPNN showed higher specificity but lower sensitivity.

These differences illustrate the effect of message-passing design on the large-scale multitask bioactivity problem. The final Tier-I representation used for downstream transfer was obtained from the selected GIN configuration, which was subsequently frozen before generation of assay-level bioactivity representations for Tier II.

C.1.2 Tier-II Backbone Comparison

We independently evaluated the same five graph architectures within the complete Tier-II multimodal framework (Table 5). In contrast to Tier I, GAT produced the strongest downstream developmental-toxicity performance. Across five held-out seeds, GAT achieved an AUROC of 0.786 ± 0.020, compared with 0.741 ± 0.039 for GIN, 0.729 ± 0.026 for MPNN, 0.727 ± 0.014 for GCN, and 0.705 ± 0.029 for AttentiveFP.

GAT was therefore used as the principal Tier-II backbone for endpoint-level analysis, gating ablations,

Table 5: Architecture sensitivity of the full Tier-II multimodal model across five matched 70/15/15 train/validation/test splits. Values are mean ± standard deviation.

Backbone	5-seed AUROC
MPNN	0.729 ± 0.026
GAT	$\mathbf{0.786 \pm 0.020}$
GCN	0.727 ± 0.014
GIN	0.741 ± 0.039
AttentiveFP	0.705 ± 0.029

and compound-specific graph interpretation.

Interestingly, the optimal backbone differed between the two learning stages. GIN performed best for the large-scale molecular multitask bioactivity problem in Tier I, whereas GAT performed best after incorporation of the structured bioactivity graph in Tier II. This result suggests that backbone performance depends on the graph-learning context rather than reflecting a universally superior architecture. In particular, the attention mechanism in GAT may be well suited to the heterogeneous, compound-conditioned neighborhood structure of the assay graph, where the relevance of assay–assay interactions varies between compounds.

C.2 Tier-I Multitask Bioactivity Architecture

Each chemical was represented as a molecular graph $G_m = (V_m, E_m)$, where atoms define nodes and chemical bonds define edges. The molecular encoder maps atom and bond features through a stack of message-passing layers followed by graph-level pooling to obtain a compound representation $\mathbf{h}_{\mathrm{mol}}$.

Five encoder families were evaluated: MPNN, GAT, GCN, GIN, and AttentiveFP. The evaluated configurations used graph-level pooling selected for each architecture; the final evaluation files include mean pooling for MPNN and GAT, attention pooling for GCN, max pooling for GIN, and mean pooling for AttentiveFP.

The final Tier-I model used the selected GIN encoder followed by a multitask classification head that simultaneously predicted activity across the retained ToxCast assays. Missing compound–assay observations were excluded from the loss rather than treated as inactive responses. Following Tier-I training, the encoder and assay-prediction head were frozen and used to generate compound-level assay activity representations for Tier II.

C.3 Tier-II Multimodal Architecture

Tier II integrates three complementary representations of each compound: (i) a molecular graph representation, (ii) a compound-conditioned bioactivity graph representation, and (iii) a physicochemical descriptor representation.

The molecular branch applies the selected graph backbone directly to the compound molecular graph. The bioactivity branch operates over the shared assay graph described in Appendix B, with assay-node states initialized from the frozen Tier-I bioactivity predictions. Learned node and edge gates adapt the contribution of individual assays and assay–assay interactions to each compound before and during message passing.

The descriptor branch transforms standardized RDKit descriptors through a multilayer perceptron. The three resulting embeddings,

$$\mathbf{z}_{\mathrm{mol}}, \qquad \mathbf{z}_{\mathrm{bio}}, \qquad \mathbf{z}_{\mathrm{desc}},$$

are concatenated,

$$\mathbf{z}_{\mathrm{fused}} = \left[\mathbf{z}_{\mathrm{mol}} \| \mathbf{z}_{\mathrm{bio}} \| \mathbf{z}_{\mathrm{desc}}\right],$$

and passed to a multilayer prediction head producing six developmental-toxicity outputs.

During training, modality-specific projection heads additionally map the three representations into a common latent space for cross-modal contrastive alignment. A masked multi-label representation objective uses only phenotype labels jointly observed for pairs of compounds, while the predictive loss excludes missing endpoint labels.

C.4 Representation-Learning Objectives

The molecular, bioactivity, and physicochemical representations are projected into a shared contrastive space using modality-specific projection heads:

$$\mathbf{p}_i^u = \frac{P_u\left(\mathbf{z}_i^u\right)}{\left\|P_u\left(\mathbf{z}_i^u\right)\right\|_2}, \qquad u \in \{\mathrm{mol}, \mathrm{bio}, \mathrm{desc}\}. \tag{11}$$

For modalities u and v, cosine similarity is

$$s_{ij}^{u,v} = \left(\mathbf{p}_i^u\right)^\top \mathbf{p}_j^v. \tag{12}$$

A symmetric InfoNCE objective aligns different representations of the same compound while separating representations belonging to different compounds:

$$\mathcal{L}_{\mathrm{CM}} = \frac{1}{3}\left(\mathcal{L}_{\mathrm{mol,bio}} + \mathcal{L}_{\mathrm{mol,desc}} + \mathcal{L}_{\mathrm{bio,desc}}\right). \tag{13}$$

For compounds i and j, jointly observed endpoints are

$$\Omega_{ij} = \left\{k : m_{ik}^Y = 1 \wedge m_{jk}^Y = 1\right\}. \tag{14}$$

Phenotypic agreement and its cosine-range target are

$$A_{ij} = \frac{1}{|\Omega_{ij}|}\sum_{k\in\Omega_{ij}} \mathbb{I}\left(y_{ik} = y_{jk}\right), \qquad T_{ij} = 2A_{ij} - 1. \tag{15}$$

For modality u, latent similarity is $C_{ij}^{(u)} = (\mathbf{p}_i^u)^\top \mathbf{p}_j^u$, giving

$$\mathcal{L}_{\mathrm{MML}}^{(u)} = \frac{\sum_{i\neq j} w_{ij}\left(C_{ij}^{(u)} - T_{ij}\right)^2}{\sum_{i\neq j} w_{ij}}, \tag{16}$$

where $w_{ij} = 1$ only when the pair satisfies the minimum shared-endpoint requirement. The implemented loss is

$$\mathcal{L}_{\mathrm{MML}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{MML}}^{(\mathrm{mol})} + \mathcal{L}_{\mathrm{MML}}^{(\mathrm{bio})}\right). \tag{17}$$

C.5 Optimization and Model Selection

Hyperparameter selection was performed using validation performance, with macro-averaged AUROC used as the principal model-selection criterion. Final Tier-II evaluation used five repeated 70/15/15 train/validation/test splits.

For the repeated holdout experiment, the five Tier-II seeds were $\{42, 184, 758, 792, 7\}$. Early stopping was based on validation AUROC, and the best validation checkpoint was restored before evaluation of the held-out test partition. Learning-rate reduction was applied when validation performance plateaued, and gradient norms were clipped during optimization.

Class imbalance and incomplete endpoint annotation were handled directly in the multitask objective. Missing labels were excluded using an observation mask, while endpoint-specific uncertainty weighting allowed the relative contribution of heterogeneous prediction tasks to be learned during training. Validation-derived classification thresholds were used only for threshold-dependent metrics; AUROC and AUPRC were calculated directly from continuous model predictions.

D Additional Experimental Analysis

D.1 Gating Ablation Study

To isolate the contribution of compound-adaptive gating within the bioactivity encoder, we performed a controlled ablation study using the GAT-based Tier-II model. The bioactivity representation consists of 457 assay nodes initialized from Tier-I bioactivity logits, with assay–assay edges described by normalized pointwise mutual information (nPMI) [Bouma, 2009] and Yule's Q [Yule, 1900].

For compound c, the full encoder applies two adaptive mechanisms. A node gate assigns a compound-specific weight to assay i from its Tier-I logit,

$$\gamma_i^{(c)} = \sigma\Big(g_{\text{node}}\left(b_i^{(c)}\right)\Big),$$

and multiplicatively modulates its node representation before message passing. An edge gate assigns a compound-specific weight to interaction (i, j) using the source and destination assay logits together with the static edge attributes,

$$\alpha_{ij}^{(c)} = \sigma\Big(g_{\text{edge}}\left[b_i^{(c)} \| b_j^{(c)} \| \text{nPMI}_{ij} \| Q_{ij}\right]\Big).$$

The resulting edge embedding is multiplicatively modulated before being passed to the GAT layers. Thus, all compounds share the same assay topology, while the relative contribution of individual assays and assay–assay interactions can vary according to the compound-specific Tier-I bioactivity profile.

We evaluated three ablated configurations against the previously evaluated full gated model: **(i) static bioactivity graph**, in which both node and edge gates were fixed to one; **(ii) node gating only**, in which node gating was retained while edge gates were fixed to one; and **(iii) edge gating only**, in which edge gating was retained while node gates were fixed to one. The **node + edge gating** condition corresponds to the complete model reported in the main experiments and was therefore not retrained separately for the ablation analysis.

Importantly, disabling a gate did not alter the underlying assay graph or its biological inputs. Tier-I assay logits, graph connectivity, and the nPMI and Yule's Q edge attributes were preserved in all conditions. The ablation therefore isolates adaptive weighting of nodes and edges rather than removing bioactivity information or changing graph topology.

Table 6: Representative compounds used for qualitative interpretation of the compound-conditioned bioactivity graph. Mechanistic concordance denotes agreement between the learned gated representation and independently reported biological activity; it does not imply causal pathway reconstruction.

Split	Compound	Phenotype profile	Known biological activity	Interpretation
Train	Mercuric chloride (527)	6/6 positive	Oxidative stress, immune signaling, and developmental injury [Zhang et al., 2016]	Broad concordance
Train	Cytarabine hydrochloride (compound 568)	0/6 positive	Incorporation into DNA and inhibition of DNA synthesis [Major et al., 1982]	Structured antimetabolite representation
Train	Cyanamide (976)	0/6 positive	Catalase-dependent activation and ALDH inhibition [DeMaster et al., 1984, 1998]	Limited direct concordance
Validation	THPC (1027)	6/6 positive	Cytotoxicity and genotoxicity [Sánchez-Fortún et al., 2005]	Broad stress concordance
Validation	Triphenyltin hydroxide (125)	Lethality only	RXR activation and endocrine disruption [Nakanishi et al., 2005]	Strong concordance
Validation	1,2,4-Trimethylbenzene (633)	0/6 positive	Predominantly neurobehavioral and central-nervous-system toxicity [U.S. Environmental Protection Agency, 2016]	Limited developmental concordance
Test	Retinoic acid (731)	6/6 positive	RAR/RXR-dependent transcription and developmental toxicity [Valcárcel et al., 1994, Oliveira et al., 2013]	Strongest held-out concordance
Test	CP-100829 (802)	Lethality only	Mitochondrial uncoupling [Hallinger et al., 2020]	Partial concordance
Test	Citric acid (313)	0/6 positive	Low ToxCast bioactivity	Negative-control example

All ablation variants retained the same GAT backbone. Specifically, the molecular encoder, RDKit descriptor branch, bioactivity-GNN dimensions, contrastive and multitask objectives, optimizer settings, and prediction head remained identical across configurations. To ensure a fair comparison, all variants were evaluated under the same 70/15/15 train–validation–test split protocol across five predefined holdout seeds ($\{42, 184, 758, 792, 7\}$).

D.2 Compound-Specific Mechanistic Interpretation Across Data Splits

To assess whether compound-adaptive gating captured chemically meaningful bioactivity patterns rather than a fixed global assay mask, we inspected the 30 highest-weighted assay–assay interactions for nine representative chemicals spanning the training, validation, and held-out test partitions (Figure 6). Collectively, the examples span broad developmental toxicity, lethality-selective toxicity, and absence of toxicity across the six grouped zebrafish endpoints. Because all compounds share the same 457-node assay topology, differences among the extracted subgraphs arise from compound-specific Tier-I bioactivity states and learned node and edge gating.

The training examples showed that the adaptive encoder learned chemically distinct graph configurations. Mercuric chloride produced a broadly distributed representation consistent with its reported oxidative-stress and immunotoxic effects, whereas cytarabine hydrochloride showed a distinct configuration consistent with its antimetabolite activity and inhibition of DNA synthesis. Cytarabine and cyanamide both retained structured representations despite being negative across all six grouped endpoints, demonstrating that high gate values are not equivalent to toxicity probabilities.

The validation compounds provided an intermediate assessment of biological coherence during model selection. In particular, triphenyltin hydroxide showed strong agreement with its established nuclear-receptor activity, including RXR agonism, whereas THPC exhibited a broader stress-related representation. 1,2,4-Trimethylbenzene provided a phenotype-negative comparison with no clear developmental receptor mechanism.

Most importantly, similar behavior was observed in the held-out test set. Retinoic acid showed the clearest mechanistic concordance: its highly weighted nuclear-receptor and transcriptional-assay relationships were consistent with RAR/RXR-mediated signaling and known retinoid developmental toxicity. CP-100829 showed partial concordance with its reported mitochondrial-uncoupling activity, while citric acid remained a structured but phenotype-negative representation. Together, these examples indicate that compound-adaptive gating can preserve biologically meaningful intermediate structure beyond the training set, while emphasizing that gated edges represent model utilization rather than direct causal pathway assignments.

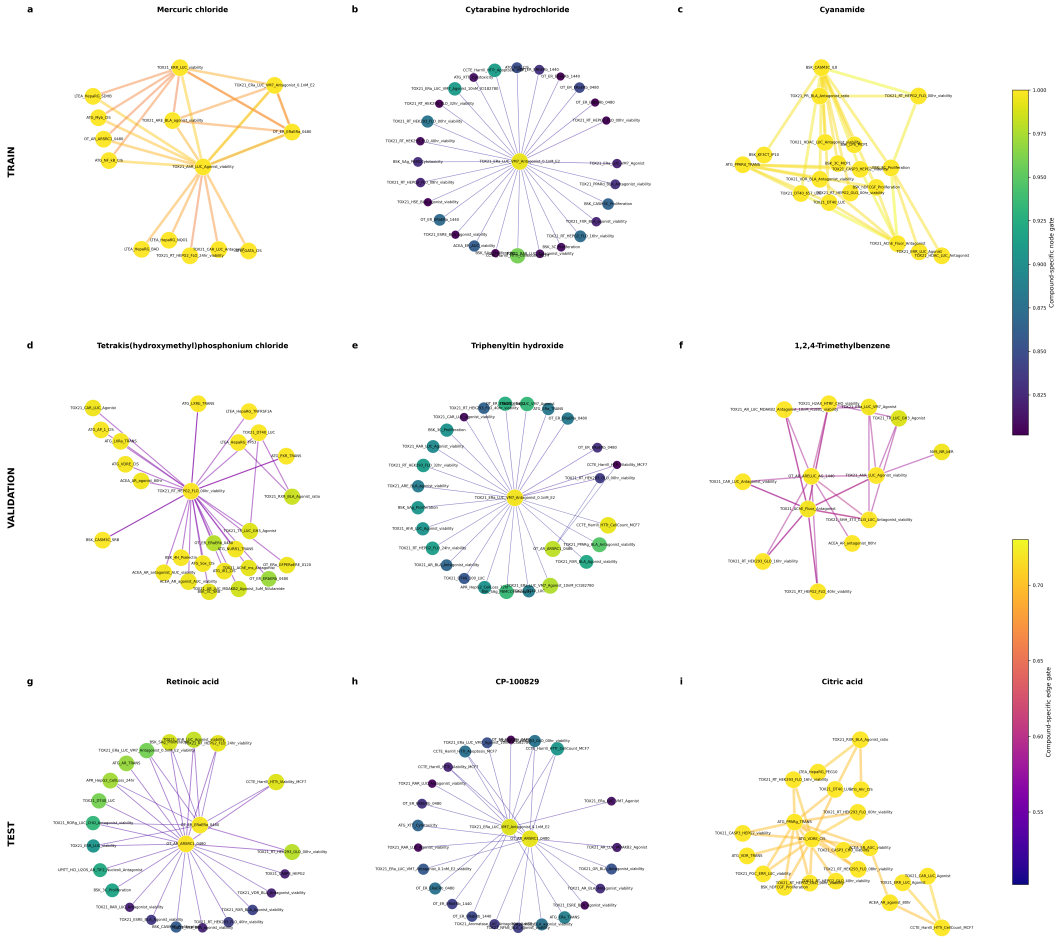


Figure 6: **Compound-specific gated bioactivity graphs across training, validation, and held-out test compounds.** The 30 highest-weighted assay–assay interactions are visualized for nine representative chemicals spanning the training (a–c), validation (d–f), and test (g–i) partitions. Training examples include (a) mercuric chloride, (b) cytarabine hydrochloride, and (c) cyanamide; validation examples include (d) tetrakis(hydroxymethyl)phosphonium chloride, (e) triphenyltin hydroxide, and (f) 1,2,4-trimethylbenzene; and held-out test examples include (g) retinoic acid, (h) CP-100829, and (i) citric acid. Nodes represent individual ToxCast assays, with node color and size indicating the compound-specific node-gate weight. Edges represent assay–assay relationships retained from the shared bioactivity graph, with edge color and width indicating the corresponding compound-specific edge-gate weight. Although all compounds share the same underlying assay topology, the selected assays and interaction patterns differ substantially between chemicals, illustrating compound-dependent utilization of the bioactivity graph. The visualizations represent model-derived assay utilization patterns and should not be interpreted as direct causal pathway reconstructions.