Patterns, Volume *3*

Supplemental information

Knowledge-guided deep learning models of drug toxicity improve interpretation

Yun Hao, Joseph D. Romano, and Jason H. Moore

SUPPLMENTARY FIGURES

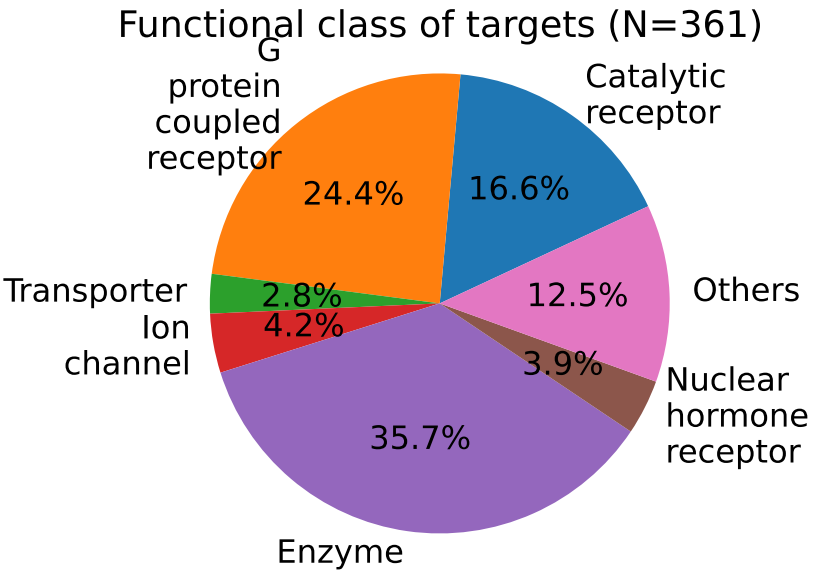

Figure S1. Distribution of functional category among proteins in the target profile

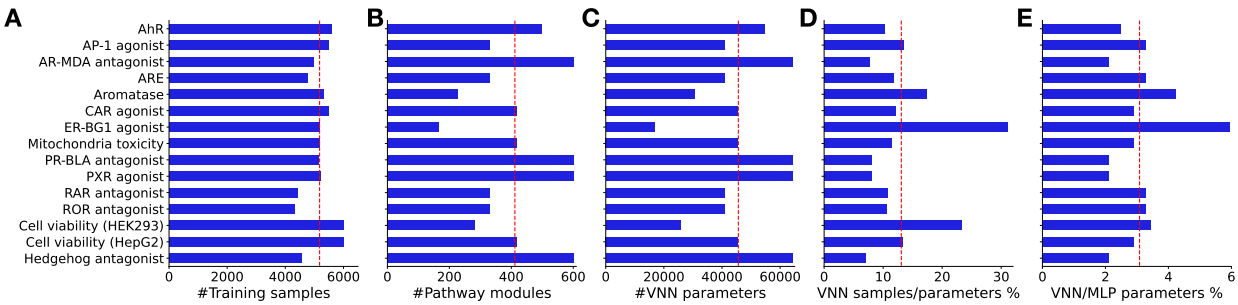

Figure S2. Comparison of DTox model statistics

Barplots showing the comparison of different model statistics for 15 toxicity assays: (**A**) the number of compounds in the training set, (**B**) the number of hidden pathway modules in the optimal DTox model, (**C**) the number of trainable parameters in the optimal DTox model, (**D**) the ratio between number of compounds in the training set versus number of trainable parameters in the optimal DTox model, and (**E**) the ratio between number of trainable parameters in the optimal DTox model versus the matched MLP model. The dashed red line in each panel represents the average across all 15 assays.

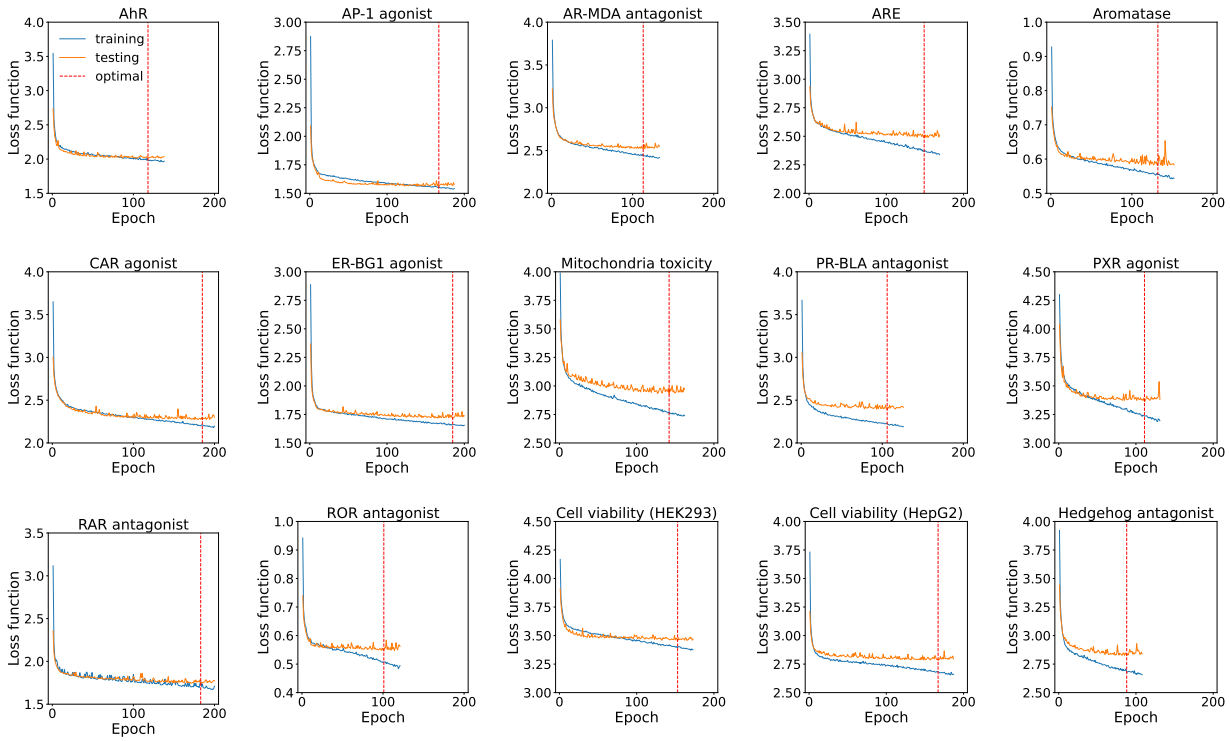

Figure S3. Evolution of loss function during learning of optimal DTox model

Line charts showing the evolution of loss function over epochs during learning process of optimal DTox models for 15 toxicity assays. Two types of loss functions are calculated and shown: loss on the training set (blue line, labeled as training) and loss on the testing set (orange line, labeled as testing). The dashed red line in each chart represents the epoch when optimal model is reached. The testing loss does not decrease for 20 consecutive epochs after the optimal point. AhR: aryl hydrocarbon receptor, AP-1: activator protein-1, AR-MDA: androgen receptor in MDA-kb2 AR-luc cell line, ARE: antioxidant response element, CAR: constitutive androstane receptor, ER-BG1: estrogen receptor in BG1 cell line, PR-BLA: progesterone receptor in PR-UAS-bla HEK293T cell line, PXR: pregnane X receptor, RAR: retinoid acid receptor, ROR: retinoid-related orphan receptor.

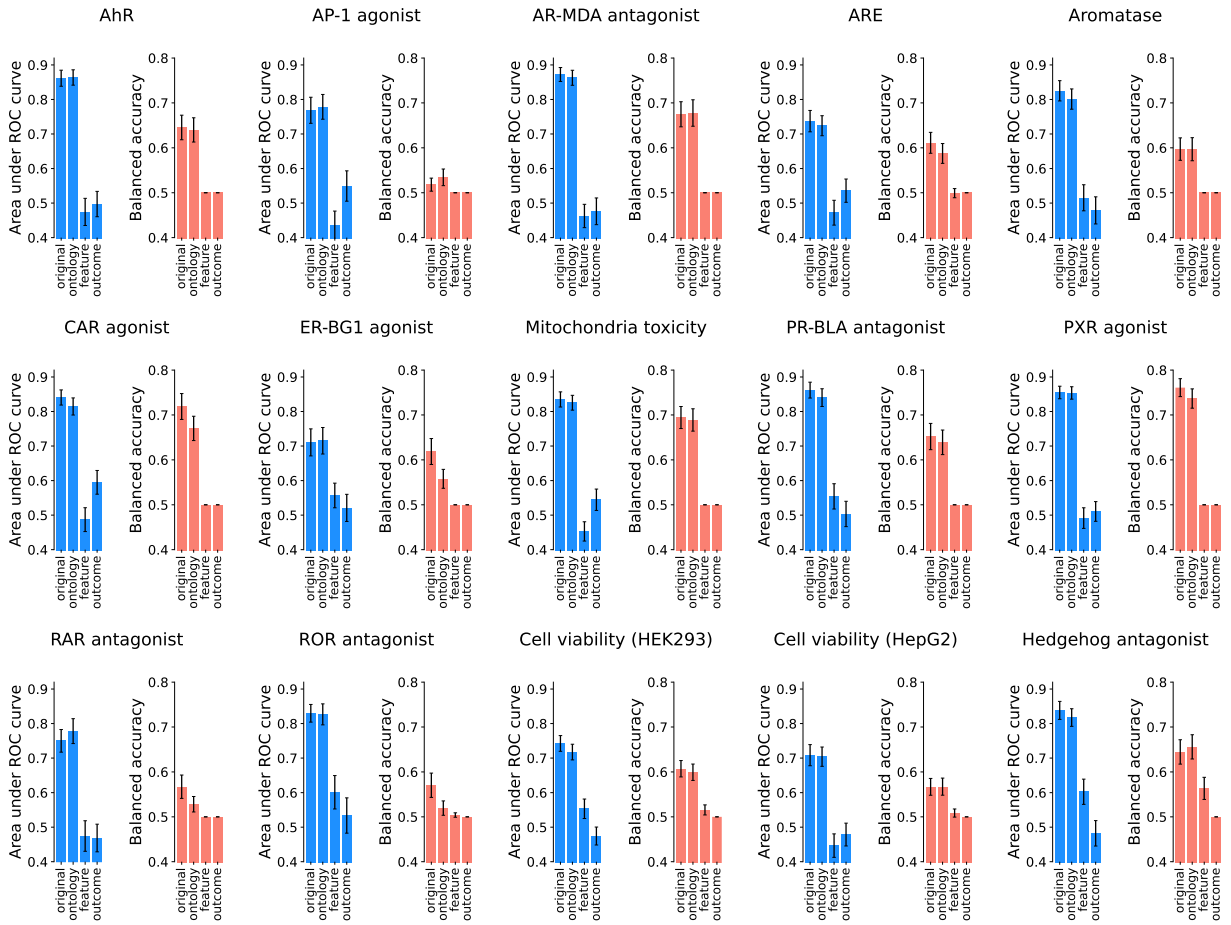

Figure S4. Influence of pathway knowledge and hierarchy on predictive performance of DTox

Barplots showing the results of shuffling analysis in all 15 Tox21 datasets. Performance of DTox (original) is compared against three alternative models built with shuffled layouts: (i) an alternative DTox model built under shuffled Reactome ontology hierarchy (ontology), (ii) an alternative DTox model built with shuffled input feature profile (feature). (iii) an alternative DTox model built with shuffled assay outcome as negative control (outcome). Performance on held-out validation set is measured by two metrics: area under ROC curve and balanced accuracy, with error bar shows the 95% confidence interval. AhR: aryl hydrocarbon receptor, AP-1: activator protein-1, AR-MDA: androgen receptor in MDA-kb2 AR-luc cell line, ARE: antioxidant response element, CAR: constitutive androstane receptor, ER-BG1: estrogen receptor in BG1 cell line, PR-BLA: progesterone receptor in PR-UAS-bla HEK293T cell line, PXR: pregnane X receptor, RAR: retinoid acid receptor, ROR: retinoid-related orphan receptor.

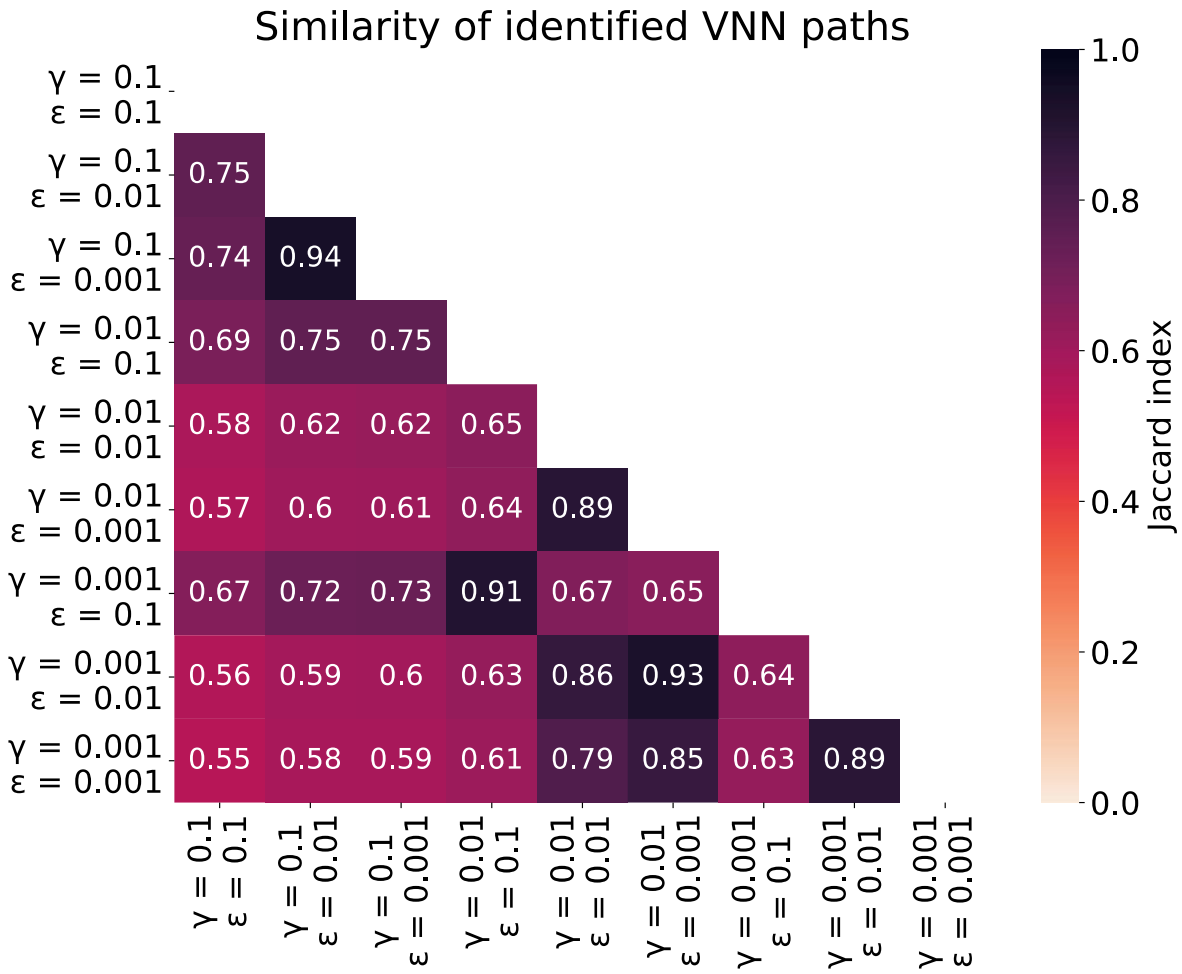

Figure S5. Consistency of DTox interpretation across hyperparameter settings

Heatmap showing the similarity of VNN paths identified from nine different hyperparameter settings. The similarity between each pair of setting (annotated in each cell) is measured by the median Jaccard Index among active compounds regarding their identified paths.

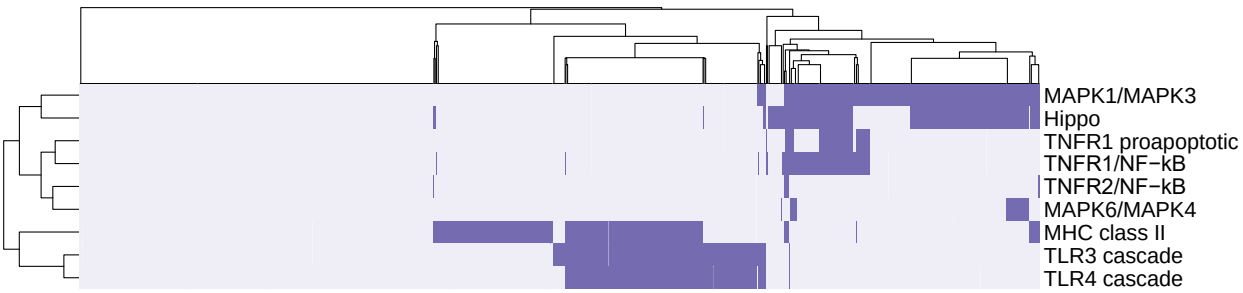

Figure S6. Clustering of HepG2-cytotoxic compounds based on cell death-related pathways

Heatmap showing the mapping between 1,120 HepG2-cytotoxic compounds (columns) and nine cell death-related pathways (rows). Hierarchical clustering is performed for both compounds and pathways. Two clusters appear to form as a result. Compounds in the first cluster (top right) are linked to cytotoxicity via apoptosis-related pathways. Compounds in the second cluster (bottom middle) are linked to cytotoxicity via immune-related and necrosis-related pathway.

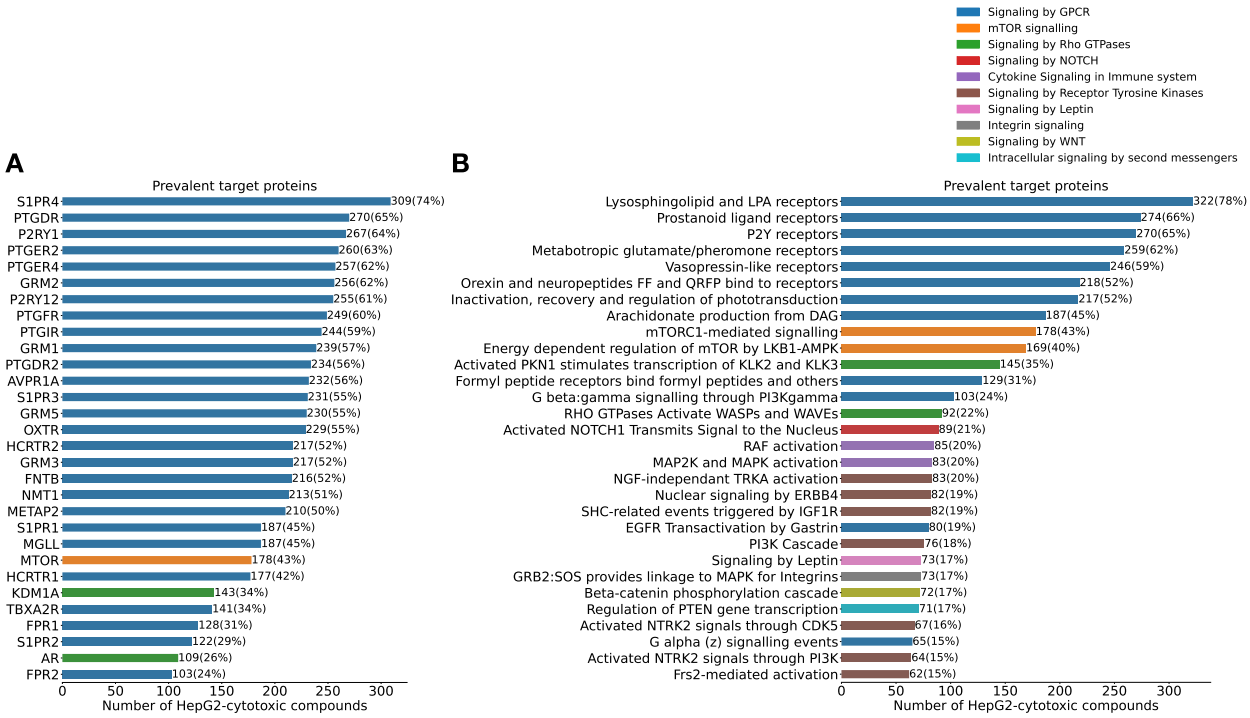

Figure S7. Summary of VNN paths identified for 413 cytotoxic compounds not mapped to cell death-related pathways

Barplots showing the top 30 most prevalent target proteins (**A**) and lowest-level pathways (**B**) identified for the 413 cytotoxic compounds that cannot be mapped to the nine cell death-related pathways. Each bar is colored by the general category of the target protein or lowest-level pathways it presents, with the color legend shown at top right.

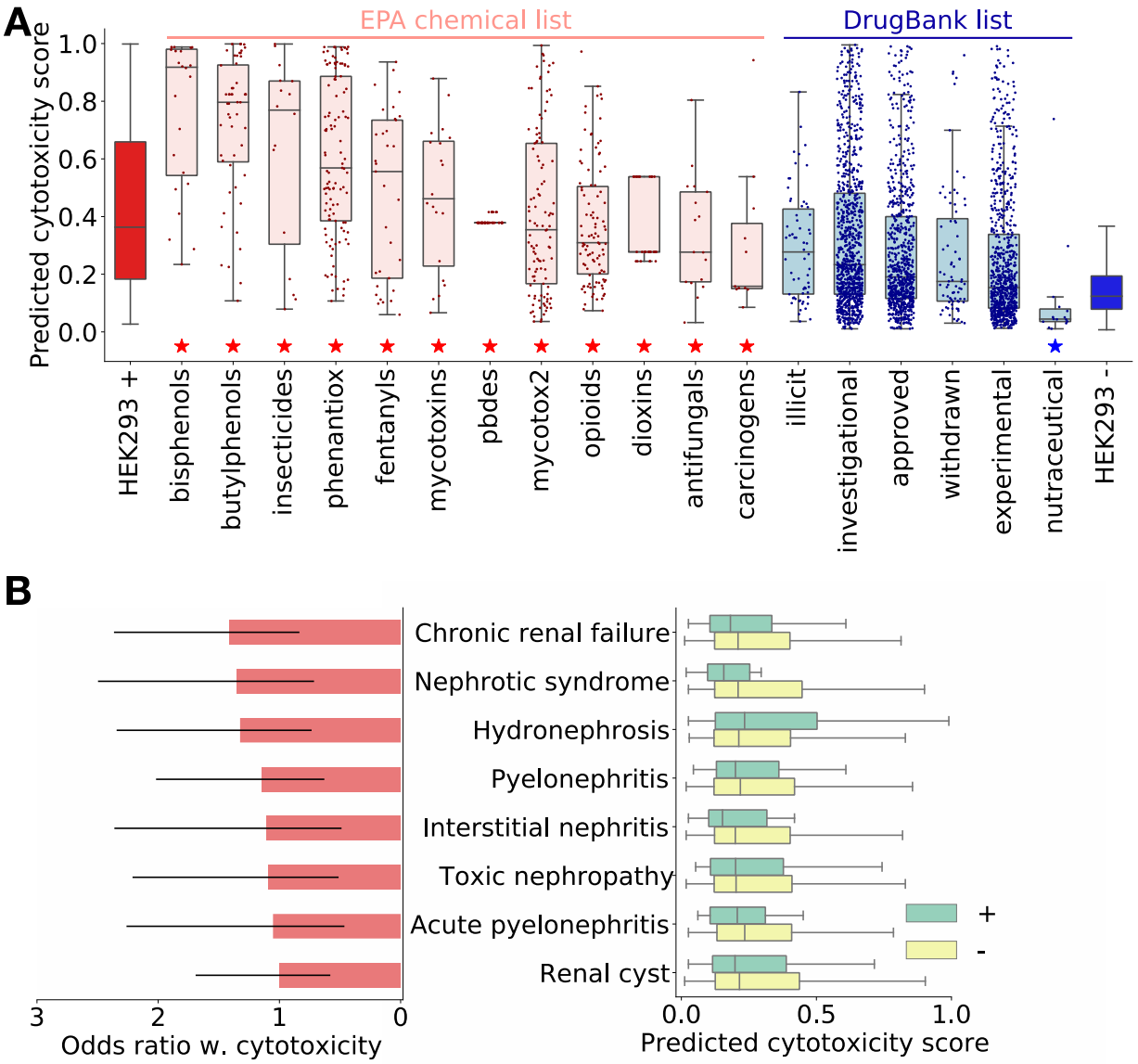

Figure S8. Application of predicted cytotoxicity score among DSSTox compounds, Related to Figure 6

(**A**) Boxplot showing the distribution of predicted HEK293 cytotoxicity scores among positive controls (leftmost box in red), 12 EPA chemical lists (boxes in light red), six DrugBank lists (boxes in light blue), and negative controls (rightmost box in blue). Mann-Whitney U test is employed to examine whether the cytotoxicity scores of each list exhibit no significant difference from the positive controls (red star above list name), or no significant difference from the negative controls (blue star above list name). (**B**) Boxplot on the right compares the predicted HEK293 cytotoxicity scores among drugs associated with clinical renal phenotypes (green box) versus negative controls (yellow box), while barplot on the left shows the odds ratio between HEK293 cytotoxicity and each phenotype (95% confidence interval shown as error bar). Results for eight phenotypes with odds ratio > 1 are shown in the plot. Mann-Whitney U test is employed to examine whether the drugs associated with each phenotype are predicted with higher cytotoxicity scores than the negative controls. No significant signal was detected for any phenotypes.

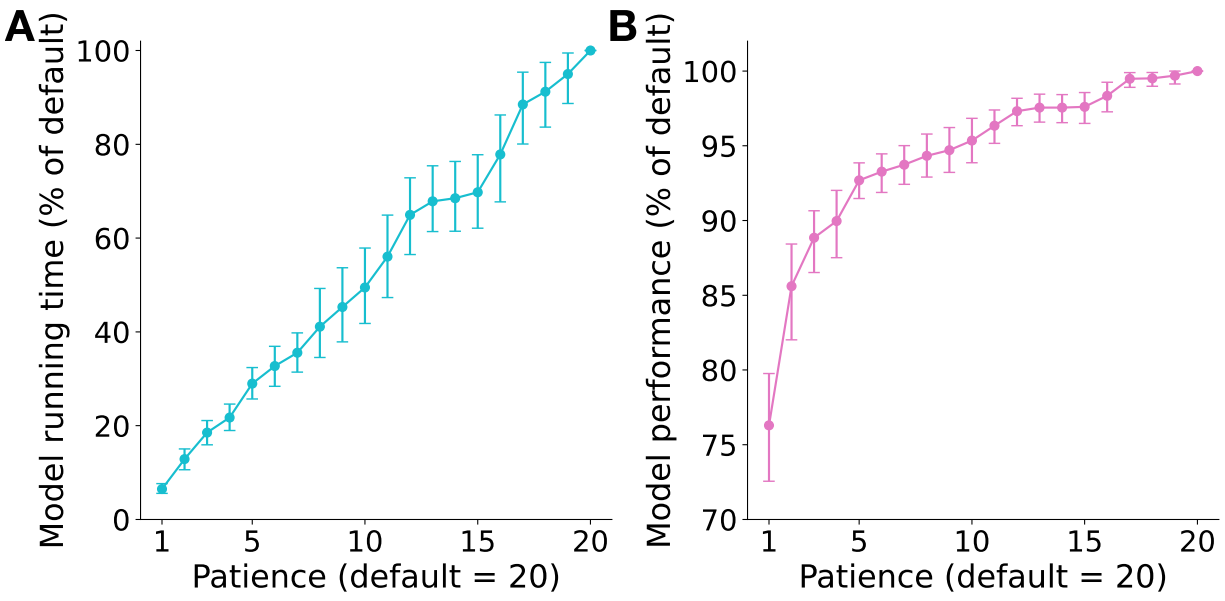

Figure S9. Influence of early stopping criterion on DTox model efficiency and performance

(**A**) Line chart showing the correlation between early stopping criterion (x-axis) and model efficiency (y-axis). The early stopping criterion is quantified by the patience hyperparameter, representing the number of epochs for which testing loss has not decreased before concluding DTox training. The efficiency is measured by relative running time, computed as the ratio between number of running epochs under alternative setting (patience = 1, 2, …, 20) versus number of running epochs under default setting (patience = 20). Each point represents the average across all 15 Tox21 datasets, with the error bar representing 95% confidence interval. (**B**) Line chart showing the correlation between early stopping criterion (x-axis) and model performance (y-axis). The performance is measured by relative improvement in testing loss (difference between optimal testing loss and testing loss after epoch one), computed as the ratio between improvement in testing loss under alternative setting (patience = 1, 2, …, 20) versus improvement in testing loss under default setting (patience = 20).