Optimization of the TeraTox assay for preclinical teratogenicity assessment

Jaklin, Manuela; Zhang, Jitao David; Schäfer, Nicole; Clemann, Nicole; Barrow, Paul; Küng, Erich

Published in:
Toxicological Sciences

DOI:
10.1093/toxsci/kfac046

Publication date:
2022

SOT | Society of Toxicology
academic.oup.com/toxsci

TOXICOLOGICAL SCIENCES, 2022, 1–17

https://doi.org/10.1093/toxsci/kfac046
Advance Access Publication Date: 29 April 2022
Research article

Optimization of the *TeraTox* Assay for Preclinical Teratogenicity Assessment

Manuela Jaklin,*,†,1,2 Jitao David Zhang,*,1 Nicole Schäfer,* Nicole Clemann,* Paul Barrow,* Erich Küng,* Lisa Sach-Peltason,* Claudia McGinnis,‡ Marcel Leist,† and Stefan Kustermann*

*Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, 4070 Basel, Switzerland; †Department for In Vitro Toxicology and Biomedicine Inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, 78464 Konstanz, Germany; and ‡Drug Discovery Unit, University of Dundee, DD1 5EH Dundee, UK

[1]To whom correspondence should be addressed. E-mail: manuela.jaklin@gmail.com and jitao_david.zhang@roche.com.

[2]Present address: Weleda AG, 4144 Arlesheim, Switzerland

Manuela Jaklin and Jitao David Zhang contributed equally to this study.

ABSTRACT

Current animal-free methods to assess teratogenicity of drugs under development still deliver high numbers of false negatives. To improve the sensitivity of human teratogenicity prediction, we characterized the *TeraTox* test, a newly developed multilineage differentiation assay using 3D human-induced pluripotent stem cells. *TeraTox* produces primary output concentration-dependent cytotoxicity and altered gene expression induced by each test compound. These data are fed into an interpretable machine-learning model to perform prediction, which relates to the concentration-dependent human teratogenicity potential of drug candidates. We applied *TeraTox* to profile 33 approved pharmaceuticals and 12 proprietary drug candidates with known *in vivo* data. Comparing *TeraTox* predictions with known human or animal toxicity, we report an accuracy of 69% (specificity: 53%, sensitivity: 79%). *TeraTox* performed better than 2 quantitative structure-activity relationship models and had a higher sensitivity than the murine embryonic stem cell test (accuracy: 58%, specificity: 76%, and sensitivity: 46%) run in the same laboratory. The overall prediction accuracy could be further improved by combining *TeraTox* and mouse embryonic stem cell test results. Furthermore, patterns of altered gene expression revealed by *TeraTox* may help grouping toxicologically similar compounds and possibly deducing common modes of action. The *TeraTox* assay and the dataset described here therefore represent a new tool and a valuable resource for drug teratogenicity assessment.

Key words: teratogenicity; molecular phenotyping; TeraTox; embryoid bodies; machine learning; factor analysis.

To assess the teratogenic potential of drug candidates, pharmaceutical companies are currently obliged to perform embryo-fetal-development (EFD) studies in at least 1 rodent and 1 nonrodent species (Beck *et al.*, 1995; ICH, 2020). There is an urgent need to develop alternative, animal-free assays for early assessment of teratogenicity. The use of humanized *in vitro* assays could potentially better mimic human physiology, reduce animal use, and lower the cost of drug development by filtering out potential teratogens early (Barrow, 2016; ICH, 2020).

Animal-free approaches based on *in silico* prediction or *in vitro* assays have been widely adopted. For instance, the CAESAR model (Cassano *et al.*, 2010) and the P&G model (Wu *et al.*, 2013) are 2 quantitative structure-activity relationship (QSAR) models for reproductive and developmental toxicity prediction. Well-established *in vitro* models for the detection of teratogenicity include the *DevTox*^{qp} assay from Stemina and the mouse embryonic stem cell test (mEST). *DevTox*^{qp} uses human-induced pluripotent stem cells (hiPSCs) to predict teratogenicity based on the ratio of ornithine and cysteine in medium supernatants (Adler *et al.*, 2008; Augustyniak *et al.*, 2019; Burridge *et al.*, 2011; Dreser *et al.*, 2020; Palmer *et al.*, 2013, 2017; Shinde *et al.*, 2015; Worley *et al.*, 2018). The mEST assay uses the beating of stem cell-derived cardiomyocytes as a functional readout for teratogenicity prediction (Genschow *et al.*, 2000, 2004; Scholz *et al.*, 1999a,b; Whitlow *et al.*, 2007).

Despite their wide adoption and relatively good performance, all these methods share major limitations. *In silico* models fail to consider the complexity and adaptiveness of biological systems. Also, their performance *a priori* for new chemical spaces that were not used for training the model is uncertain. On the other hand, existing *in vitro* models, which rely on a single readout for prediction, have limited capability to probe the complex biological processes underlying drug-induced teratogenicity.

To overcome these limitations, we developed a new, humanized *in vitro* teratogenicity assay for the preselection of pharmaceutical candidates. The new assay, which we call *TeraTox*, uses ethically nonrestricted hiPSC-derived embryoid bodies (EBs) that differentiate spontaneously into all 3 germ layers, with expression of representative early developmental markers of each layer. The new assay gave promising results in a preliminary evaluation (Jaklin *et al.*, 2020). This article describes the subsequent characterization and critical assessment of *TeraTox*, including a detailed predictive algorithm. We tested a panel of 45 pharmaceuticals with evidence of human teratogenicity profiles based on FDA classification using a 6-point concentration-response, generating the largest dataset so far in a single study about *in vitro* modeling of pharmaceutical teratogenicity with reference to clinical or animal data. We adapted an amplicon-based RNA sequencing technique (a technology known as *Molecular Phenotyping*) to quantify the expression of germ-layer genes as well as pathway reporter genes. We benchmarked a variety of architectures of machine-learning models and identified the best-performing predictive model using factor analysis and random-forest regression. *TeraTox* performed favorably compared with other models, ie, the prediction accuracy was comparable with that of both Stemina *DevTox*^{qp} and mEST assays (higher sensitivity, lower specificity) and higher than that of 2 QSAR models (higher sensitivity and specificity). More importantly, *TeraTox* offers insights into a multitude of changes caused by the compounds on gene, pathway, and germ-layer levels, some of which corroborated their teratogenicity potential.

MATERIALS AND METHODS

hiPSCs-derived TeraTox *assay*. The *TeraTox* assay is built upon a commercially available hiPSC line (Gibco, A18945), which has indistinguishable gene expression profiles compared with embryonic stem cells (Burridge *et al.*, 2011; Quintanilla *et al.*, 2014). The cells form 3D EBs and undergo multilineage differentiation into all 3 germ layers (Jaklin *et al.*, 2020). Prior to the assay, the hiPSCs were tested with the TaqMan ScoreCard assay (Thermo Fisher) to confirm sufficient levels of pluripotency (Tsankov *et al.*, 2015a). The EBs were spontaneously differentiated and treated with each reference substance over a time course of 7 days in Elplasia 96w micro-well plates (Corning, 4442) using the ViaFlo 96 automated microplate pipetting device (Integra) for liquid handling. Compounds were applied to the EBs on days 0, 3, and 5 at 6 concentrations, together with EB medium and 0.25% DMSO solvent controls as the negative reference. To test for batch-to-batch variation, we included several positive reference compounds in multiple runs (eg, hydroxyurea, valproic acid, SB431542, etc.). Cell viability was determined prior to gene expression studies on day 7 by measuring ATP release in supernatants with the CellTiter-Glo 3D assay (Promega, G9681) according to the manufacturer's protocol to prespecify appropriate testing ranges. Concentrations that showed <80% cell viability were excluded from the subsequent gene expression studies. CellTiter-Glo reagent (100 µl) was added and incubated for 5 min on a shaker to lyse the EBs. The plates were kept for an additional 25 min in the dark at room temperature for binding of the released ATP to the luminescent dye. ATP release in supernatants was measured with the spectrophotometer (Biotek, Vermont). All cell culture media and reagents were obtained from Gibco (Thermo Fisher) unless otherwise specified. The overall cell culture and cytotoxicity protocols have been described previously in detail by Jaklin *et al.* (2020). Targeted gene expression profiling was performed in biological duplicates at 6 subcytotoxic concentrations using the molecular phenotyping platform described previously (Drawnel *et al.*, 2017; Zhang *et al.*, 2014, 2015). The resulting 1055 samples of differentiated EBs were lysed after 7 days in 350 µl MagNA Pure LC RNA Buffer (Roche Diagnostics) and purified using an automated MagNA Pure 96 system (Roche Diagnostics). The total RNA was quantified using the Qubit RNA Assay Kit (Thermo Fisher) on the Fluorometer Glomax (Promega). Total RNA, with a maximum of 10 ng from each biological replicate, was reverse transcribed to cDNA using Superscript IV Vilo (Thermo Fisher). Libraries were generated with the AmpliSeq Library Plus Kit (Illumina) according to the reference guide. Pipetting steps for target amplification, primer digestion, and adapter ligation were done with a miniature mosquito automatic pipettor (SPT Labtech). For the purifications before and after final library amplification, solid-phase reversible immobilization magnetic bead purification (Clean NGS, LABGENE Scientific SA) was performed on a multidrop automated pipetting station (Thermo Fisher). We measured both amplicon sizes and cDNA concentrations using an Agilent High Sensitivity DNA Kit (Agilent Technologies) according to the manufacturer's recommendation. Prior to sequencing, cDNA contents of the samples were normalized and pooled to 2 nM final concentration on a Biomek FXP workstation. The libraries were sequenced on the NovaSeq 6000 Instrument (Illumina) using sequencing-by-synthesis technology. All 75 cycles ended up with a minimum of 2 Mio sequencing reads per sample for analysis. We used molecular phenotyping with 1215 detectable pathway reporter genes, including a subset of 87 early developmental markers (germ-layer genes, Supplementary Table 4), and genes representative of toxicological pathways to identify differentially expressed genes induced at prespecified concentration levels (Tsankov *et al.*, 2015a,b).

Assessing characteristics of differentiated hiPSCs with BioQC. We applied the *BioQC* software developed previously to characterize the identity of the differentiated samples across all treated compound concentrations (including vehicle controls) on day 7

(Zhang *et al.*, 2017). We used raw data of gene expression derived from molecular phenotyping and compared these profiles with tissue-preferential gene signatures derived from organ, tissue, and cell-type-specific gene expression data compiled from public compendia (Ljosa *et al.*, 2013; Young *et al.*, 2008). The BioQC performs Wilcoxon-Mann-Whitney tests comparing expression of genes in a set, eg, genes preferentially expressed in 1 tissue, versus genes that are not in the set. The enrichment scores (log-10 transformed *p* values) reported by BioQC are used to assess the similarity between the expression profile of interest and cell-type- and tissue-specific expression profiles.

Analysis and modeling of the TeraTox *data*. We performed differential gene expression (DGE) analysis comparing compound-treated samples with DMSO controls using the generalized linear model implemented in the *edgeR* package in R/Bioconductor (Robinson *et al.*, 2010). To generate features for machine-learning models, we transformed the *p* values associated with the coefficients of compound treatment to z-scores by the inverse of the quantile function of Gaussian distribution, multiplied by the sign of $\log_2$ fold-change (logFC). The vectors of z-scores of all genes ($n = 1215$) were used as raw features for machine-learning models, based on which further feature selection and engineering work were performed. We also tested the possibility of using the effect size, logFC, as a feature.

Besides the raw feature set of z-scores of all genes, we used 3 knowledge- and data-driven approaches to engineer the features in order to improve the performance of the machine-learning algorithms. First, we confined ourselves to the subset of germ-layer genes, because our and other's work confirmed that their expression is specific to germ layers of embryogenesis, and their expression is modulated by teratogenic compounds (Supplementary Table 4; Bock *et al.*, 2011; Jaklin *et al.*, 2020; Tsankov *et al.*, 2015a,b). Second, we used the germ-layer associations reported by Tsankov *et al.* to derive a reduced feature set defined by 5 germ-layer classes, including both germ layers (ectoderm, endoderm, mesoderm, and mesendoderm) and pluripotency, by taking the median z-scores of germ-layer genes associated with each germ-layer class (Tsankov *et al.*, 2015a). Finally, we used factor analysis, a dimension-reduction approach that derives latent variables from the correlation structure of observed variables, to identify latent biological, germ-layer factors (germ-layer factors for short), which reflect linear combinations of transcription factors, epigenetics, and other gene regulatory mechanisms that control embryogenesis.

We predicted teratogenicity potential in 2 ways. One way was to treat teratogenicity as a binary variable and to perform binary classification. The other way was to convert concentration-response teratogenicity into numeric metrics and to construct regression models. For the latter case, we defined a compound-specific Teratogenicity Score (TS hereafter). For nonteratogens, the TS was defined as 0 independent of the tested concentration. For teratogens, the TS was defined as the 0-1-bounded cosine similarity between the differential expression profile induced by a given concentration of a compound and the differential expression profile induced by the highest noncytotoxic concentration of the same compound. The noncytotoxic concentration was determined as the highest tested concentration associated with an average viability equal or larger than 80%.

The models were trained and validated using the Leave-One-Out (LOO) scheme. The full panel of compounds was assessed successively, leaving out 1 compound at a time and then used to build machine-learning models. We then compared TS predicted by the models with the observation of each left-out compound using the Spearman correlation coefficient. As an alternative to LOO, we also assessed repeated 80%/20% splitting of data into training sets and test sets.

In short, we considered 2 types of features (z-scores and logFC), 4 sets of factors (all genes/germ-layer genes/median z-scores or logFC of germ-layer classes defined by Tsankov *et al.*/median z-scores or logFC of germ-layer factors defined by factor analysis), 2 methods (linear regression with elastic net regularization and random forest, implemented in the *caret* package, version 6.0-88), 2 types of target variables (binary classification and regression), and 2 training/testing schemes (LOO and 80%/20% splitting). We tested all combinations exhaustively to build machine-learning models for TS and identified the best-performing models.

Besides predicting TS, we also comprehensively probed all options to build regression models for cytotoxicity (100%-viability), which was measured as part of the *TeraTox* assay. The same set of model architectures was tested; however, the combinations giving best performing models differed from that for TS (further discussed in Results section). All data analysis was performed with R (version 4.0.1) or Python (version 3.8.1) unless otherwise specified.

Test chemicals for validation. In total, we tested 28 positive and 17 negative reference substances in 6-point concentrations in the mEST (see Supplementary Material and Methods, Supplementary Figs. 1a, 1b, and Supplementary Table 1) and the human *TeraTox* assay (Table 1). This compound panel consisted of both commercial and developmental pharmaceuticals with known teratogenicity potential (ie, positive or negative) available from either human data, as reported in FDA drug labels, or from *in vivo* EFD studies in rats and/or rabbits (ICH, 2020). Some compounds without existing human or *in vivo* animal data were classified as teratogens based on a known teratogenic hazard associated with their mode of action (Belair *et al.*, 2020; Chen *et al.*, 2002; Cusack *et al.*, 2017; Evans, 2007; Kameoka *et al.*, 2014; Lipinski *et al.*, 2008; Sakata and Chen, 2011; Wang *et al.*, 2013; Worley *et al.*, 2018). Compounds that did not result in increased incidences of birth defects in an adequate prospective cohort study accepted by health authorities were considered as nonteratogenic in humans, at least at the therapeutically relevant exposure levels (Adams *et al.*, 1969; Daniel *et al.*, 2019; Dashe and Gilstrap, 1997; Etwel *et al.*, 2014; Muanda *et al.*, 2017; Rumbold *et al.*, 2015; Supplementary Table 2).

The commercial compounds were obtained from Merck, Germany. The 12 investigational small molecule drug candidates RO-1 to RO-12 were provided by F. Hoffmann—La Roche, Switzerland (compound structures are not disclosed due to confidentiality and intellectual property issues). No human pregnancy data were available for the investigational drug candidates, but *in vivo* data were available from EFD studies in rats, and/or in rabbits (Supplementary Table 3). RO-1, RO-3, RO-8, RO-9, and RO-10 were teratogenic in EFD studies; RO-2, RO-4, RO-5, RO-6, RO-7, RO-11, and RO-12 did not induce teratogenicity.

All compounds were serially diluted in DMSO (0.25%) from a stock solution to 6 test concentrations and tested at appropriate noncytotoxic concentration ranges in the *TeraTox* and *mEST* assays. We used the following metrics to compare the performance of the *TeraTox* and *mEST* assays. Sensitivity was calculated as the proportion of correctly predicted teratogens. Assay specificity was calculated as the proportion of correctly predicted nonteratogens. Overall accuracy was taken as the proportion of all correct predictions. F_1 scores were calculated as the harmonic mean of precision and recall. True positive, true

Table 1. Reference Compounds in the Human *TeraTox* Assay

Reference Compound	Teratogenicity Classification	Test Concentrations (Human Model) [μM]	CAS Number
Acitretin	**Positive**	0.08–2.5 (1:2)	55079-83-9
Amoxicillin	Negative	6.25–200 (1:2)	26787-78-0
Artesunate	**Positive**	0.13–8 (1:2)	88495-63-0
Ascorbic Acid	Negative	28–900 (1:2)	62624-30-0
Bosentan	**Positive**	4.7–150 (1:2)	147536-97-8
Busulfan	**Positive**	0.06–4 (1:2)	55-98-1
Carbamazepine	**Positive**	4.7–300 (1:2)	298-46-4
Cetirizine	Negative	9.3–600 (1:2)	83881-51-0
Cyclopamine	**Positive**	0.3–20 (1:2)	4449-51-8
Cyproheptadine	Negative	0.47–30 (1:2)	129-03-3
Dabrafenib	**Positive**	0.03–2 (1:2)	1195765-45-7
DAPT	**Positive**	0.05–3 (1:2)	208255-80-5
Dasatinib	**Positive**	0.3–20 (1:2)	302962-49-8
Dexamethasone	**Positive**	4.7–300 (1:2)	50-02-2
Dorsomorphin	**Positive**	0.2–14 (1:2)	866405-64-3
Doxycycline	Negative	0.3–20 (1:2)	564-25-0
5-Fluorouracil	**Positive**	0.004–0.25 (1:2)	51-21-8
Hydroxyurea	**Positive**	3.12–200 (1:2)	127-07-1
Ibuprofen	Negative	1.9–1400 (1:3)	15687-27-1
Isotretinoin	**Positive**	4.7–300 (1:2)	4759-48-2
Imatinib	**Positive**	1.6–100 (1:2)	152459-95-5
IWP-2	**Positive**	0.0015–0.1 (1:2)	686770-61-6
Lazabemide	Negative	1.6–100 (1:2)	103878-84-8
Metformin	Negative	7.8–500 (1:2)	657-24-9
Methotrexate	**Positive**	0.0025–40 (1:5)	59-05-2
Misoprostol	**Positive**	0.02–1.3 (1:2)	59122-46-2
Penicillin G	Negative	9.3–600 (1:2)	61-33-6
Progesterone	Negative	0.63–40 (1:2)	57-83-0
Retinoic Acid	**Positive**	0.0005–0.035 (1:2)	302-79-4
RO-1*	**Positive**	1.6–100 (1:2)	n/a
RO-2*	Negative	7.8–500 (1:2)	n/a
RO-3*	**Positive**	4.7–300 (1:2)	n/a
RO-4*	Negative	3.1–200 (1:2)	n/a
RO-5*	Negative	0.8–50 (1:2)	n/a
RO-6*	Negative	6.25–400 (1:2)	n/a
RO-7*	Negative	9.3–600 (1:2)	n/a
RO-8*	**Positive**	1:25–80 (1:2)	n/a
RO-9*	**Positive**	0.08–5 (1:2)	n/a
RO-10*	**Positive**	0.23–15 (1:2)	n/a
RO-11*	Negative	0.6–40 (1:2)	n/a
RO-12*	Negative	1.6–100 (1:2)	n/a
SB431542	**Positive**	0.31–20 (1:2)	301836-41-9
(±) Thalidomide	**Positive**	0.0007–0.5 (1:3)	50-35-1
Valproic Acid	**Positive**	15.6–1000 (1:2)	99-66-1
Warfarin	**Positive**	0.9–60 (1:2)	81-81-2

Compounds for assay validation, with human teratogenicity classification and test concentration according to noncytotoxic concentrations. Dilution ratios in brackets covering 6 concentrations. Teratogenicity classification was based on FDA classification (Supplementary Table 2) or *in vivo* EFD data (indicated with asterisks*, Supplementary Table 3).
Abbreviation: EFD, embryo-fetal-development.

negative, false positive, and false negative are denoted with TP, TN, FP, and FN, respectively, and the performance metrics are defined in Supplementary equations 1–5. To identify the threshold of TS that maximizes the performance (F_1 score) of the *TeraTox Score*, we used a grid search between 0 and 1 with a step size of 0.01. The best threshold (TS = 0.38) was chosen manually by inspecting the performance metrics.

To benchmark the performance of *TeraTox*, we applied 2 regulatory-accepted structure-based models to predict teratogenicity of commercially available compounds: the CAESAR model (version 2.1.8, Cassano *et al.*, 2010) and the P&G model (version 1.1.2, Wu *et al.*, 2013) implemented in the VEGA platform (version 1.3.10, Marzo *et al.*, 2016). For the benchmark, we used 20 compounds (15 teratogens and 5 nonteratogens) that were not part of the training set of the CAESAR model.

Model explainability and interpretation. We used the type I importance measure of features (mean decrease in accuracy) of random-forest models to compare the importance of germ-layer genes in the teratogenicity model and in the cytotoxicity model.

Pharmacology data of publicly available compounds were downloaded from ChEMBL (version 26). We only used human

targets and affinities derived from high-quality dose-response data. Binary distances were used to cluster the compounds by their pharmacological profiles.

To construct a Bayesian network model of regulations between factors, we first discretized DGE data of the first 6 germ-layer factors into 3 levels using the Hartemink's pairwise mutual information method implemented in the *bnlearn* package (Scutari, 2010). We generated 1000 bootstrap replicates using Hill Climbing, a score-based learning algorithm, and the Bayesian Dirichlet equivalent (uniform) score (bde, with the imaginary sample size set to 10). Edges that persisted in more than 85% bootstrap samples were deemed as significant and reported.

The beta regression model used for sensitivity analysis was built with the *glmmTMB* package (Brooks *et al.*, 2017). Scores outside the boundaries [0.01, 0.99] are set to the boundary values to allow beta regression. All 10 factors and significant interaction terms identified in the Bayesian network were used as the model input and compounds were modeled as random effects to capture between-concentration correlations. For better interpretability, input variables were scaled to zero mean and SD. Simulation was performed using the *ggeffects* package (Lüdecke, 2018).

RESULTS

Gene Expression Quantification by Molecular Phenotyping

We previously described that differential expression of a set of 87 genes preferentially expressed in different germ layers is in principle able to distinguish between teratogenic and nonteratogenic compounds (Jaklin *et al.*, 2020). These germ-layer genes both determine and reflect embryonic development (Tsankov *et al.*, 2015a). To validate our findings, we compiled a large set of well-documented teratogens and nonteratogens that are challenging to predict and/or known to cause FP in animal studies (Supplementary Tables 2 and 3). The compounds cover a broad spectrum of chemical classes and a wide range of effective concentrations.

We evaluated the performance of our human stem-cell model by testing the panel of compounds, adapting the experimental workflow developed previously (Figs. 1A and 1B). We identified the assay throughput as a major challenge due to the high number of samples for gene expression profiling (>1000). It would be particularly cost- and labor-intensive to use the digital PCR technique established in our previous work to quantify gene expression (Jaklin *et al.*, 2020). To address this challenge, we used molecular phenotyping as an alternative readout. Molecular phenotyping is based on amplicon-based targeted sequencing and is able to deliver quantitative expression data of 1215 predefined genes, including both pathway reporter genes, ie, genes that are specifically modulated by pathway perturbations, as well as germ-layer genes that we reported in our previous study. In this way, we were able to characterize both general pathway activity modulations and germ layer-specific changes as potential features associated with teratogenicity (Drawnel *et al.*, 2017; Zhang *et al.*, 2014, 2015).

We performed extensive quality control of the data. Here we address the questions whether results of molecular phenotyping are comparable with those of qRT-PCR, and whether the hiPSCs used show expected cell identity based on their gene expression profile. We compared the differential expression profiles of germ-layer genes obtained by qRT-PCR in previous studies with newly generated data of molecular phenotyping and observed highly consistent results (Pearson correlation coefficient $R = 0.9$, $p < 2.2E16$; Figure 1C).

Molecular phenotyping requires far fewer cells and delivers much higher throughput than qRT-PCR, allowing a marked improvement in the productivity of the *TeraTox* assay. A unique advantage of quantifying pathway reporter genes along with germ-layer genes is the identification of cell-type-specific gene expression patterns. To this end, we applied *BioQC* analysis, a method that we developed to identify sample heterogeneity and tissue comparability using gene sets preferentially expressed in cells and tissues (Zhang *et al.*, 2017). We observed that the expression profiles of the cells used in the *TeraTox* assay at day 7 resemble a mix of those gene signatures specific for astrocytes, epithelial cells, and iPSC-derived neurons (Figure 1D). This suggests that the hiPSCs used for the assay has a preferred differentiation propensity into the neuroectodermal lineage, which is in agreement with previous time-series gene expression studies that demonstrated pronounced expression of ectodermal markers at day 7, followed by meso- and endodermal expression (Jaklin *et al.*, 2020; Tsankov *et al.*, 2015a).

Unsupervised Learning From Gene Expression Data With Factor Analysis

Before applying supervised learning techniques to differentiate teratogens from nonteratogens, we applied several unsupervised learning algorithms to explore the gene expression data, including principal component analysis (PCA) and factor analysis. PCA revealed experimental plate effects that we could successfully correct with linear regression models for DGE (data not shown). Unexpectedly, factor analysis revealed both biological insights and suggested a technique for feature engineering to produce the best-performing model (see below, also a brief introduction to factor analysis is given in Supplementary Materials and Methods).

We applied factor analysis to raw gene expression data and identified intriguing patterns. Since factor analysis is based on intergene correlations, we visualized the correlation matrix of germ-layer genes in Figure 2A (the full matrix is visualized in Supplementary Figure 2a). Genes that strongly correlate with each other form clusters, which correspond to latent factors. Despite that factor analysis is a correlation-based statistical method in which we injected no prior knowledge, biologically meaningful patterns emerged. Using the maximum likelihood method, we decomposed the covariance matrix of gene expression into factors. The heatmap in Figure 2B shows loadings, ie, how strong factors influence the expression of germ-layer genes, of the first 10 factors that collectively explain more than 70% of the covariance (Supplementary Figs. 2b–d). Left to the heatmap we use colors to indicate germ-layer classes that were distilled from biological knowledge. We found that the first 6 factors (ranked by explained covariance of the data) are significantly enriched with signatures of individual germ layers or signatures of stem-cell self-renewal (Figure 2C, $p < .01$, Fisher's exact test).

This significant enrichment is intriguing, because while it is established that germ-layer genes are highly expressed at different stages of embryogenesis, factor analysis reveals for the first time that their expressions are strongly correlated in 3D EBs formed by hiPSCs, with or without compound treatment. Given that the cells used in *TeraTox* are cultured up to day 7, it is unlikely that the correlations are caused by temporal changes of embryogenesis. Instead, factor analysis suggests that besides being correlated across time in development, expression of

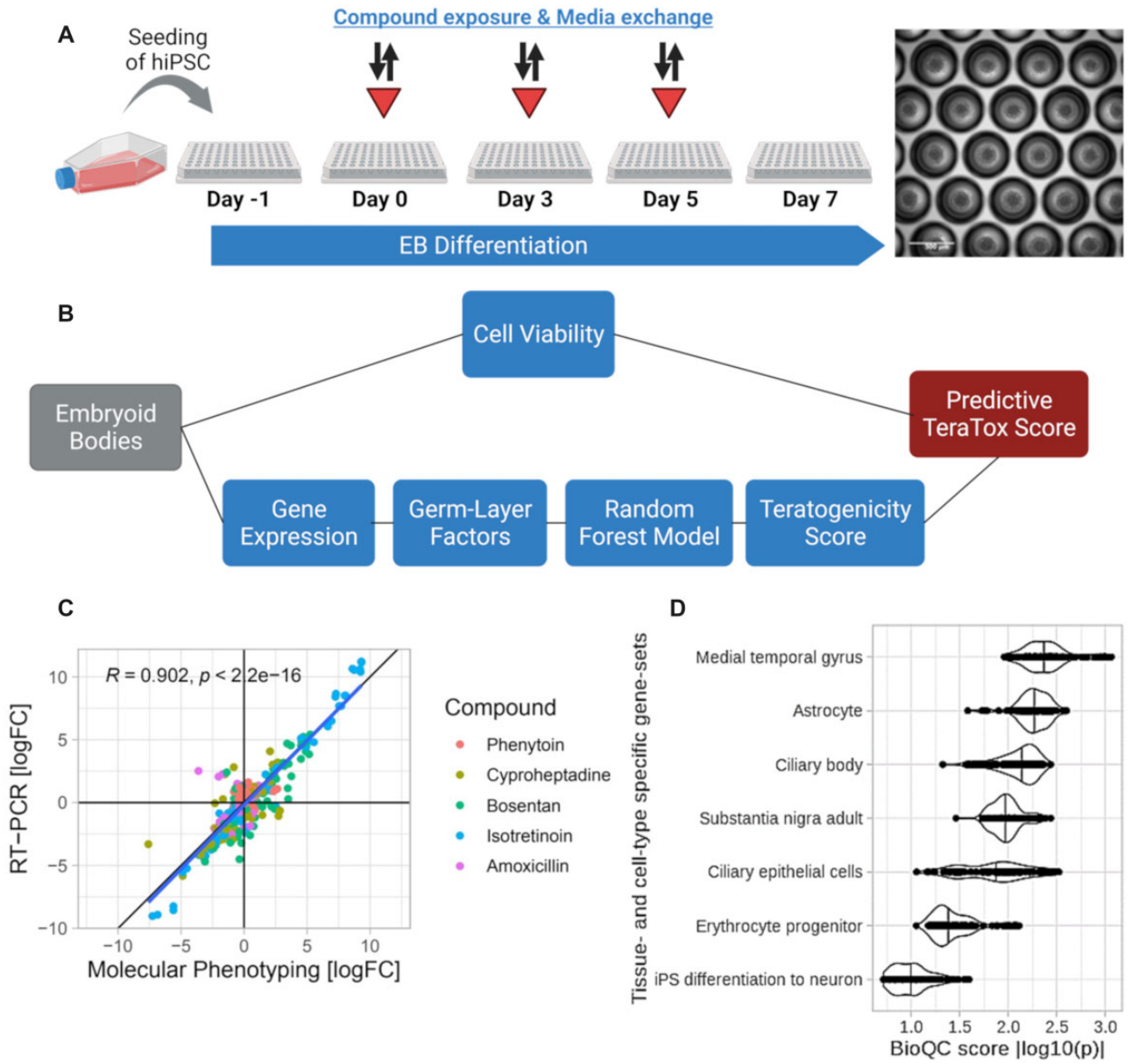

Figure 1. The human *TeraTox* assay: workflow and quality control. A, hiPSCs differentiate over 7 days and form EBs. Compounds are added on days 0, 3, and 5. Single wells of differentiated EBs are lysed for 1 sample. B, *TeraTox* score is calculated based on cell viability, gene expression data, and machine-learning model. C, DGE from molecular phenotyping correlated with data from RT-PCR assay represented in log2 fold change (log2FC). Dots represent germ-layer genes. R = Spearman correlation coefficient between 2 sets of measurements in all compounds. D, *BioQC* of raw gene expression data (DMSO controls) revealed the biological identity of hiPSCs and showed significantly enriched cell-type signatures (median $p < .10$). Each dot = 1 sample. Violins show distributions of BioQC scores (absolute log10 transformed p values of the Wilcoxon-Mann-Whitney test) from each gene set, vertical lines indicate median values. The larger the BioQC score, the more enriched is the expression of the genes. Abbreviations: hiPSCs, human-induced pluripotent stem cells; EBs, embryoid bodies; DGE, differential gene expression.

germ-layer genes is also correlated across treatment conditions in 7-day spontaneously differentiated EBs.

Detailed analysis of the results from the factor analysis revealed more insights. The strongest correlation of the germ-layer genes was observed among genes in Factor 1, many of which are markers of the ectodermal layer, eg, *WNT1*, *POU4F1*, *OLFM3*, *CDH9*, *LMX1A*, *DMBX1*, *PAX3*, *MAP2*, and *TRPM8* (Figure 2A). Although BioQC analysis revealed that ectodermal genes are highly expressed at the endpoint on day 7, factor analysis further indicated that their expression is strongly correlated across conditions, too, which is neither sufficient nor necessary for their high expression. Factors 2–6 mainly consist of genes representing the mesodermal layer (factor 2), stem-cell self-renewal (factor 3), and the endoderm layer (factors 4–6), respectively. The remaining factors (factors 7–10) are of smaller sizes and more heterogeneous (Figure 2B). Genes associated with each factor are associated mainly, but not exclusively, with other genes of the same germ-layer class. In summary, factor analysis revealed that germ-layer genes form coregulated gene modules in *TeraTox* that are significantly enriched by germ-layer- or stem-cell-specific markers.

Training and Testing of a Predictive Model for the TeraTox *Assay*

To build a quantitative predictive model of concentration-dependent teratogenicity potential with gene expression, we explored all combinations of the following options (Figure 3A):

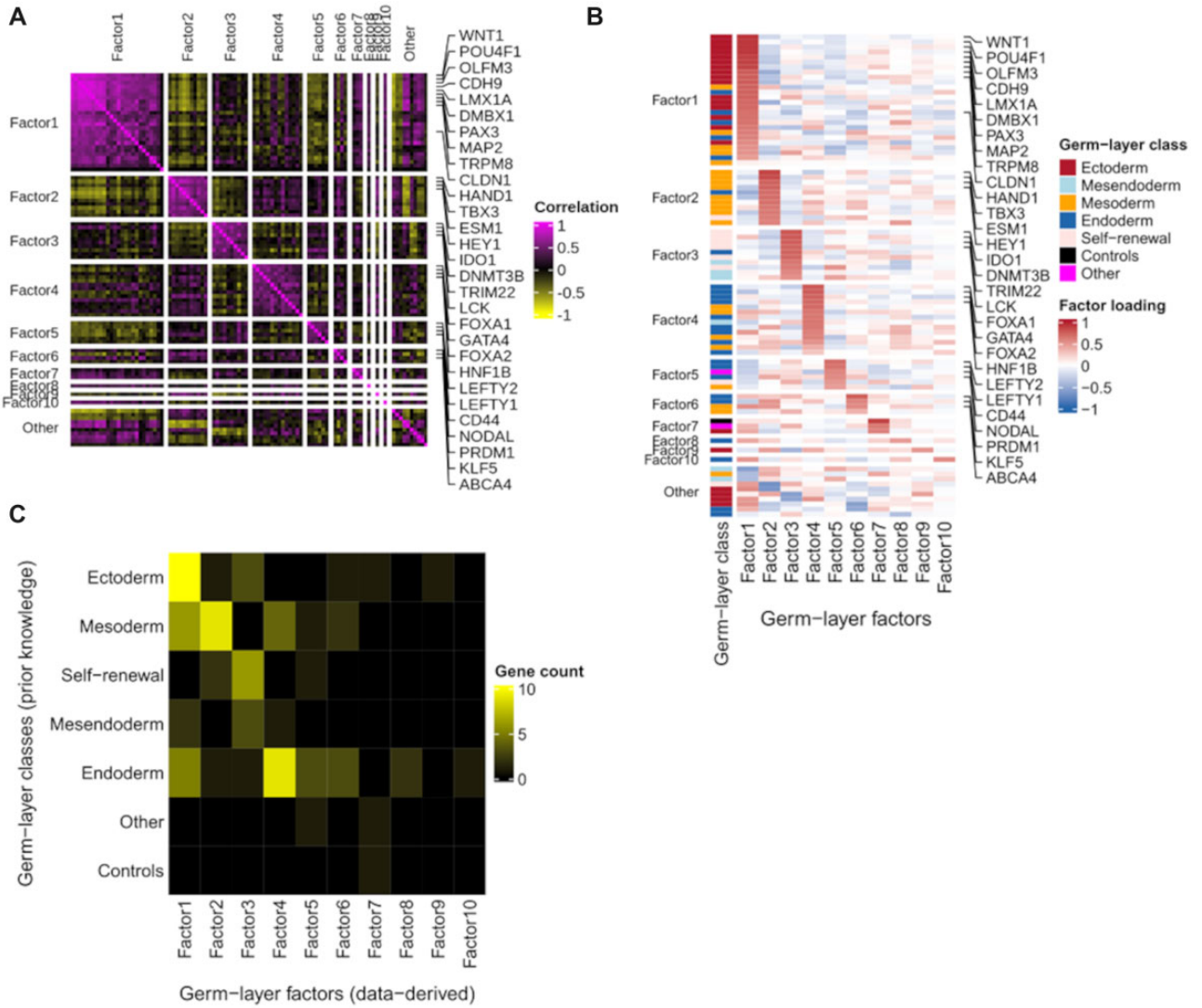

Figure 2. Identification of latent factors that are associated with germ layers. A, Germ-layer genes show correlated expression and form clusters of coexpression. The heatmap represents pairwise Pearson correlation coefficients of germ-layer gene expression in all samples. Genes are split by latent germ-layer factors for representative genes (full matrix is shown in Supplementary Figure 2a). B, Loadings of factor analysis (germ-layer genes in rows and linear combinations of latent germ-layer factors in columns). Loadings equal to or near +1 or −1 indicate that the factor positively or negatively influences the gene, while loadings near 0 means that the factor has little effect on the gene (full matrix in Supplementary Figure 2c). C, Germ-layer factors are not equivalent to, but significantly associated with, germ-layer classes. The heatmap visualizes the number of genes shared by each pair of germ-layer classes (in rows) and germ-layer factors (in columns).

1. *Feature type*: We tested both logFC, the point-estimate of the effect size, and z-scores transformed from the sign of logFC and p value reported by the *edgeR* model, which considers both effect size and variance of DGE.
2. *Feature engineering*: We used all detectable pathway reporter genes ($n = 1215$), detectable germ-layer genes ($n = 87$), germ-layer classes defined by Tsankov *et al.* ($n = 7$), and germ-layer factors derived from factor analysis ($n = 10$). For both germ-layer classes and factors, we used the median value of the genes belonging to each group as the engineered feature.
3. *Model construction*: We used and benchmarked 2 methods of different nature, Elastic Net (linear regression with regularization) and Random Forest (ensemble decision trees), to construct machine-learning models. These methods were chosen based on the size of the dataset and the relatively good explainability of both methods (Badillo *et al.*, 2020).
4. *Target variable*: We used both binary classification (teratogen or nonteratogen) and regression (the TS, defined below and further detailed in the Materials and Methods section) for teratogenicity and regression alone for cytotoxicity.
5. *Data splitting*: we tried both repeated splitting of 80% training and 20% test set, and the LOO scheme. For data splitting, we used 80% of compounds (stratified sampling from nonteratogens and teratogens) as the training set to train a model, which was used to predict the TS using the remaining 20% compounds as the test set. For LOO, the model was trained by assessing the panel of compounds minus one, which predicted the TS for the left-out compound. The procedure was repeated until all compounds had been left out. The performance of both models was assessed by F_1 scores in case of binary classification models, and Spearman correlation coefficients of TS for teratogens in case of regression models.

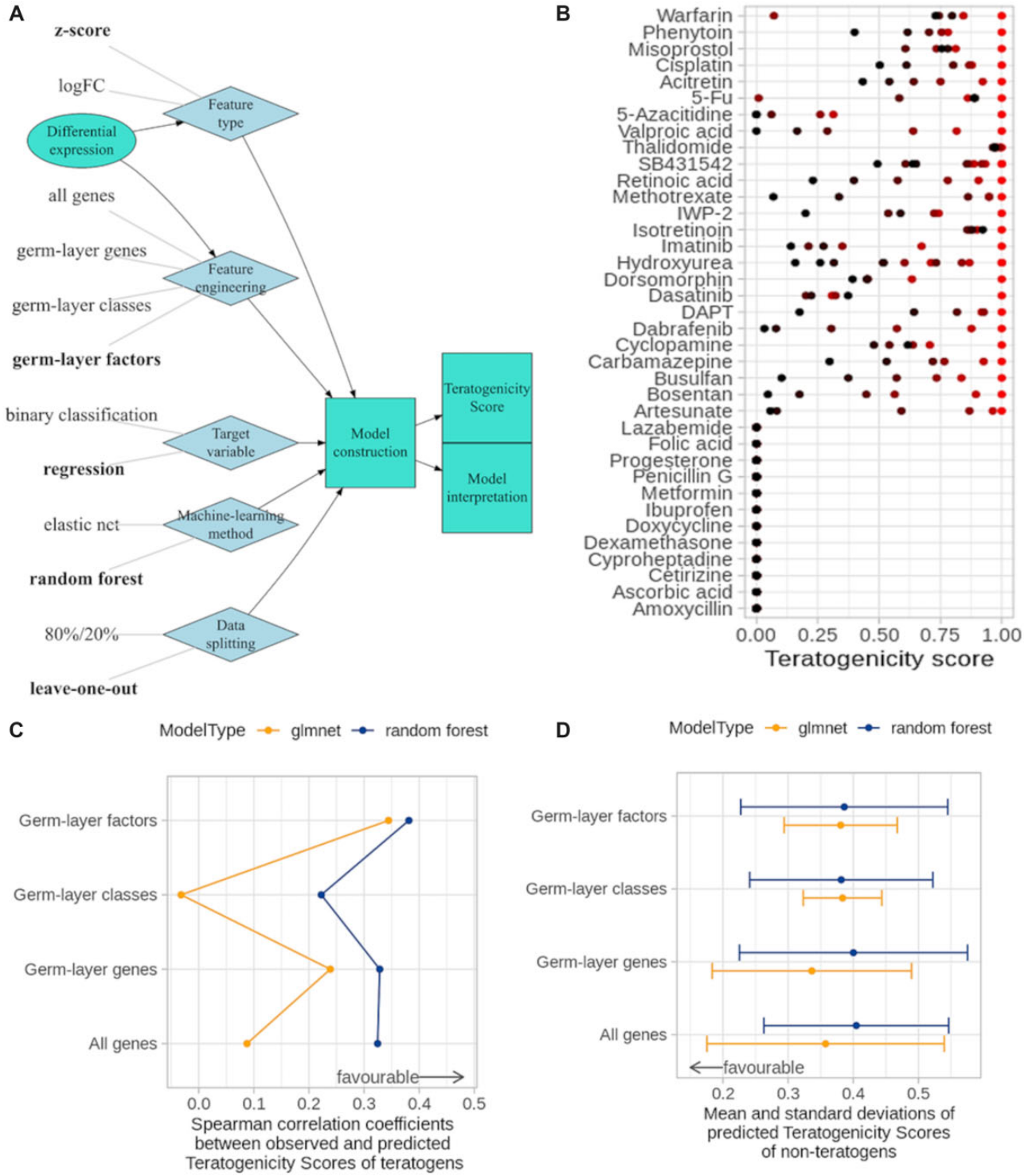

Figure 3. Construction of machine-learning models predicting concentration-dependent teratogenicity potentials based on differential gene expression as input. A, Overview of investigated model architectures. B, Definition of TS from lowest (left) to highest (right) concentrations. TS of teratogens = 1 for the highest noncytotoxic concentration, TS for other concentrations = cosine similarity of differential gene expression profiles between each concentration and the highest noncytotoxic concentration. TS of nonteratogens = 0, independent of the concentration level (from *Lazabemide* to *Amoxicillin*). Negative TS = 0. C, Spearman correlation coefficients between observed teratogenicity scores, calculated on a per-compound basis, and predicted teratogenicity scores, which are derived from models trained by LOO testing. D, Mean (dots) and SDs (error bars) of teratogenicity scores of nonteratogens. Median teratogenicity score of each compound is derived from 6 concentrations. The average scores of nonteratogens are lower than those of teratogens, but not strictly zero, because they are predicted values by LOO testing with our machine learning model instead of assigned values as in (B). Abbreviations: TS, Teratogenicity Score; LOO, leave-one-out.

The best model parameters were searched by 10-fold cross-validations of the training set.

The TSs of teratogens are defined between 0 and 1, and those of nonteratogens are fixed as 0 at all concentrations (Figure 3B). By defining TS, we effectively transformed the binary classification problem into a regression problem. We refer readers interested in the motivation of developing the TS and in the mathematical details to the section on Teratogenicity Score in Supplementary Materials and Methods.

We observed the following patterns as we tried all options of model building:

1. The *feature type* has minimal impact on the performance, though models trained with *z*-scores perform better on the test set than models trained with logFC (data not shown).
2. The combination of *feature engineering* and *machine-learning model* is important and the best combination depends on the prediction task (Figs. 3C and 3D). For teratogenicity prediction, the combination of germ-layer factors and random-forest regression worked best.
3. With regard to the *target variable*, the performance of the regression-based teratogenicity-score prediction model is slightly better than binary classification (data not shown).
4. Performance is comparable between 2 modes of *data splitting* (data not shown). However, the LOO training-testing scheme is preferable because it allows us to set up a single threshold of TS, which can be applied to all compounds, and is not conditioned by whether or not a compound is included in the training set or in the test set as in the case of 80%/20% data splitting.

Based on these observations, we decided to use germ-layer factors as features, random-forest regression as the machine-learning model, and TS as the target variable to build the predictive model for teratogenicity with gene expression data.

Performance of the TeraTox *Assay and Benchmarking With Other Models*

Based on the best-performing machine-learning model, we defined the following predictive model for teratogenicity. First, we considered the maximal noncytotoxic threshold concentration (NCC_{max}) for cell viability of at least 80%, measured by the CellTiter Glo assay. Next, we defined the minimal teratogenic concentration (TC_{min}) as the concentration at which the threshold of the TS was met (TS = 0.38, defined by grid search; Figure 4A). If no NCC_{max} or TC_{min} could be determined because values did not exceed these thresholds, the maximal tested concentrations were used for NCC_{max} and TC_{min}. The predictive score, which we named *TeraTox Score,* is defined by the logarithmic ratio between threshold concentrations at 20% viability impairment (NCC_{max}) and teratogenic concentrations (TC_{min}). Negative *TeraTox* scores classify the compounds as negative whereas positive scores classify compounds as positive (Figure 4B).

We plotted the concentration-response curves of measured cytotoxicity and predicted TS induced by each compound (Figure 4C; see Supplementary Figure 4 for all compounds). In general, teratogenicity levels increased while cell viability decreased with rising concentrations. Correctly predicted negative compounds were unlikely to induce teratogenicity within noncytotoxic concentrations, which means the calculated *TeraTox* score was negative or zero (eg, Doxycycline, RO-4, RO-6). Positive compounds (eg, Bosentan, Carbamazepine, Retinoic Acid, RO-1) or FP predicted compounds (eg, Cetirizine) were more likely to induce teratogenicity under noncytotoxic concentrations, as indicated by positive *TeraTox* scores (Figure 4C).

We compared the *TeraTox* prediction scores with classifications from FDA or *in vivo* EFD studies for 45 reference compounds (Supplementary Table 5). Classification with *TeraTox* Scores achieved an overall accuracy of 69% and outperformed mEST (58%). The 2 assays show different sensitivity and specificity profiles: Although mEST is more specific (specificity 76%), *TeraTox* is more sensitive (sensitivity/recall 79%). Among 17 negative reference compounds, 8 were classified as FP by *TeraTox*, and only 4 by the mEST. Whereas from 28 positive reference compounds, 22 were predicted as TP by *TeraTox* and only 13 by the mEST (Table 2 and Figure 4D). It is noteworthy that among the 26 compounds misclassified in total, the following 7 compounds are wrongly predicted by both assays: cyproheptadine, RO-11, 5-FU, methotrexate, misoprostol, RO-8, and warfarin. Given the distinct sensitivity and specificity profiles of the 2 assays, we asked whether we can achieve even better prediction results by using the 2 tests in sequence. Therefore, if we first run the mEST on the full panel, the substances with negative mEST results would then be retested by *TeraTox* to benefit from the high specificity of mEST and the high sensitivity of *TeraTox*.

Indeed, we found that overall accuracy of the combined prediction increased to 78%, better than either *TeraTox* or *mEST* alone. This suggests that it may be possible to achieve better prediction results by combining the existing mEST assay with the novel *TeraTox* assay.

Furthermore, we compared 18 pharmaceutical compounds that were both tested in *TeraTox* and *DevTox*^{qp} by Stemina, and observed the identical accuracy (78%), whereas balanced accuracy was 73% for *TeraTox* and 87% for *DevTox*^{qp} assay. The *DevTox*^{qp} assay delivered a higher specificity (100%) compared with *TeraTox* (67%), whereas *TeraTox* was more sensitive (80%) than *DevTox*^{qp} (73%; Supplementary Tables 6 and 7).

We also compared *TeraTox* with in *silico* predictions of developmental and reproductive toxicity using 2 widely used QSAR models: CAESAR and P&G, both implemented in the VEGA software. Among the compounds that we tested, 20 compounds are new to the CAESAR model, the P&G model, and the LOO versions of the *TeraTox* model, namely these compounds were not used to train these models and therefore, the prediction results for them present a fair comparison of the performance (Supplementary Table 8). Among these 20 compounds, *TeraTox* performed better (85% accuracy) than both the CAESAR model (75% accuracy) and the P&G model (35% accuracy; Supplementary Table 9).

In short, *TeraTox* delivers comparable or more favorable performance with alternative assays, and combining *TeraTox* with other assays can further increase the prediction accuracy for drug-induced teratogenicity.

Leveraging TeraTox *Data and Model to Gain Biological Insights*

A model's interpretability and explainability is crucial to allow for inspection and further improvement (Barredo Arrieta *et al.*, 2020). We performed additional in-depth analysis of the cytotoxicity and gene expression data and collected additional data orthogonal to *TeraTox*, thereby implementing 4 independent approaches to interpret and explain how the machine-learning model works and to explore what differs teratogens from nonteratogens.

First, we followed up on previous work and asked the question whether compound-induced cytotoxicity quantified by the phenotypic assay can be predicted by gene expression data as well, and whether TS are confounded by general cytotoxicity

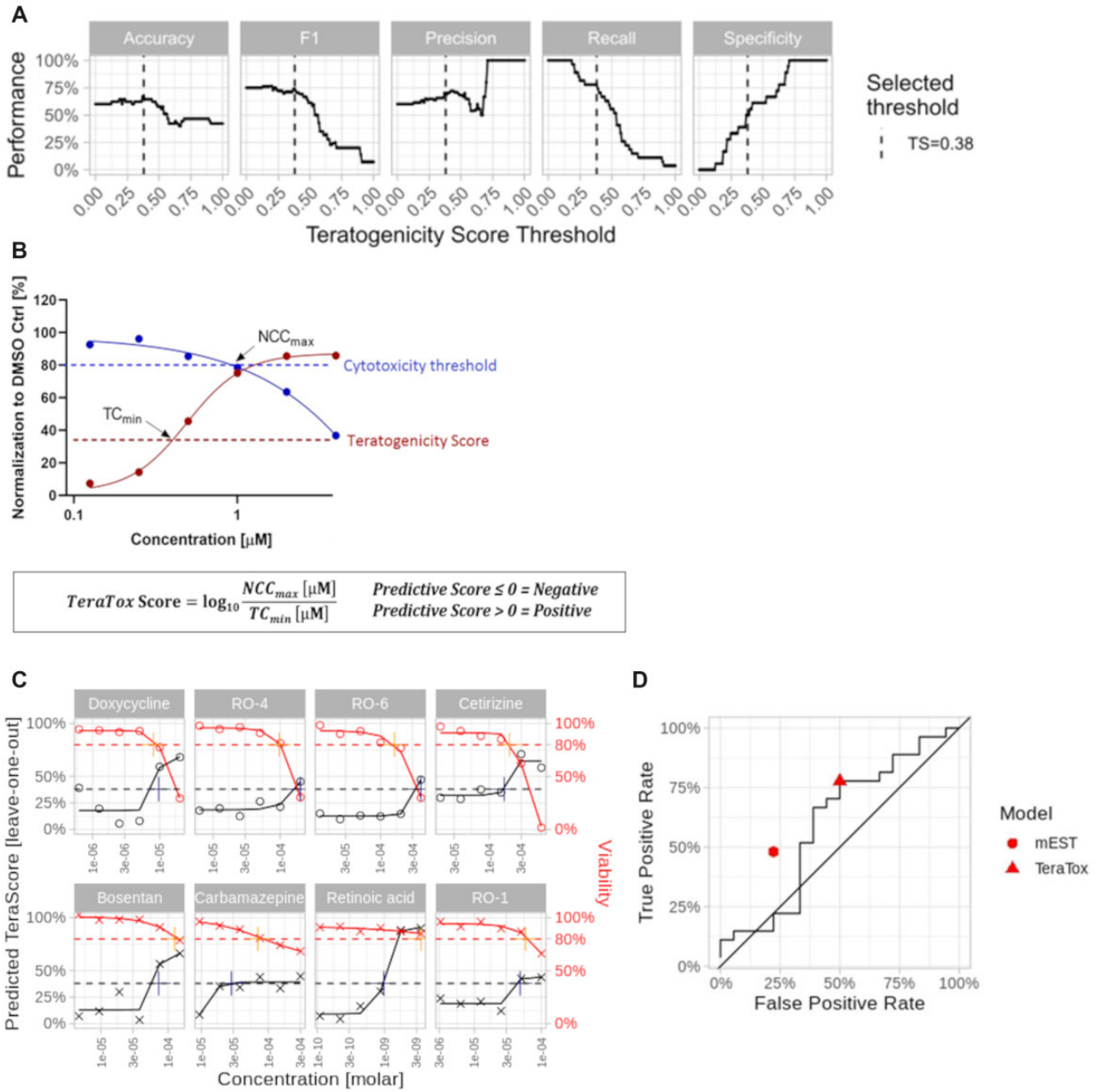

Figure 4. Prediction of teratogenicity with the human TeraTox assay. A, Results of a grid search to select the optimal threshold of the TS using the best identified model architecture. The best threshold (TS = 0.38) was chosen based on performance metrics defined in Supplementary equations 1–5. B, Visual definition of the TeraTox score based on minimal teratogenic concentration (TC_{min}) and maximal noncytotoxic concentration (NCC_{max}). Compounds with TeraTox scores ≤0 are classified as negative and compounds with scores >0 are classified as positive. C, Examples of concentration-response curves reported by the TeraTox assay of 4 selected nonteratogens (top panels) and 4 selected teratogens (bottom panels). The "+" indicates predicted TS ($n = 2$) and measured cytotoxicity ($n = 3$). D, ROS curve of LOO tests based on 45 reference compounds. Abbreviations: TS, Teratogenicity Score; ROC, receiver operating characteristics; LOO, leaving-one-out.

Table 2. Overview of Assay Performance for mEST and Human *TeraTox* Assay

Model	TP	TN	FP	FN	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1 (%)
TeraTox	22	9	8	6	69	73	79	53	76
mEST	13	13	4	15	58	76	46	76	57
mEST + TeraTox	22	13	4	6	78	85	79	76	82

Values were calculated based on 45 compounds (according to Supplementary equations 1–5).
Abbreviations: mEST, mouse embryonic stem cell test; TP, true positive; TN, true negative; FP, false positive; FN, false negative.

(Waldmann *et al.*, 2014, 2017). For this purpose, we followed the same scheme as described in Figure 3A using cytotoxicity instead of TS as the target variable. Interestingly, a comprehensive search showed that using all pathway reporter genes and the *elastic net* model, instead of using germ-layer factors and random forest as in the case of teratogenicity prediction, give the best result (Figure 5A, contrasted with Figs. 3C and 3D).

Given that the combination of germ-layer genes and random forest gives reasonable performance in both cases, and that random forest allows inquiry of feature importance by accuracy, we compared the feature importance of germ-layer genes in predicting both target variables (Figure 5B). The prediction of cytotoxicity and teratogenicity by molecular phenotyping relies on expression changes of distinct genes. The distinction shows that (1) teratogenicity of a compound is not a determinant for cytotoxicity, (2) a compound that shows cytotoxicity at a specific concentration can still be teratogenic at lower concentrations, and (3) genes and pathways associated with cytotoxicity and teratogenicity can be regulated independently from each other. This concurs with several previous findings (Krug *et al.*, 2013; Rempel *et al.*, 2015; Shinde *et al.*, 2017).

The second approach addressed the question whether a compound's pharmacology, in this context its target profile (protein targets and binding affinities), suffices to predict its teratogenicity potential. If so, one may hope to predict teratogenicity potential based on target profiles and/or even based on the chemical structure alone. Although some teratogens indeed have similar target profiles, we observe close clustering of teratogens and nonteratogens that have similar target profiles as well (Figure 5C and Supplementary Figure 5a). The potential of teratogenicity, therefore, may be associated with off-target effects or effects through targets that are not captured in ChEMBL, especially at the relatively high concentrations approaching cytotoxicity levels that we tested. Corroborating this, we found almost no correspondence between clustering of average DGE across concentration per compound and that of pharmacological profile (Supplementary Figure 5b). Therefore, we conclude that while knowing the target- and off-target profile of a compound is essential for de-risking its safety liabilities including teratogenicity, pharmacology data alone cannot currently predict a compound's teratogenicity potential. This conclusion concurs with the superior performance of *TeraTox* over 2 QSAR models, which consider the chemical structure alone. The hiPSCs-based *in-vitro* assays, such as *TeraTox* and other advanced cellular models, are therefore indispensable for assessing the potential for human teratogenicity.

The third approach was to use a simpler generalized linear regression model for sensitivity analysis, which would allow us to analyze how the model responds to changes of the input. Given that random forest is an ensemble method and the contribution of each germ-layer factor can be therefore difficult to interpret, we built an alternative model using generalized linear regression. To identify interaction terms in the linear regression, we made the assumption that germ-layer factors regulate each other by forming a directed acyclic graph. Under this assumption, we built a Bayesian network using the differential expression data of germ-layer factors (Figure 5D). The network reveals potential influences on both mesoderm and endoderm by the ectoderm, influences on endoderm by mesoderm, and influences on stem-cell renewal by endoderm.

The Bayesian network topology prompted us to build a beta regression model including all germ-layer factors and interactions identified in the Bayesian network (Figure 5E and Supplementary Figure 6). The model provides both interpretable coefficients of the model and a tool for sensitivity analysis, because we can quantify prediction uncertainty more easily with a generalized linear model than with a random forest model, by paying the price of assuming linear regulation relationship. For the sensitivity analysis, we kept all other parameters fixed and tuned one input parameter at a time to simulate its impact on predicted TSs. We observed that the model is likely sensitive to impairment of either ectoderm layer or stem-cell self-renewal, while being relatively robust to changes to either mesoderm or endoderm (Figure 5E). The results of sensitivity analysis further underlined the prominent ectodermal nature of the model at the endpoint on day 7.

Last but not least, we applied gene-set enrichment analysis to each compound with *BioQC* and compared median gene-set enrichment results over concentrations of each compound between teratogens and nonteratogens. We identified multiple gene-sets that are potentially differentially regulated by teratogens and nonteratogens ($p < .10$; Supplementary Figure 7). Interestingly, target genes of several transcription factors that are involved in organ development and cell differentiation, eg, POU4F1, GATA1, and NODAL, were impacted by teratogens differently from nonteratogens. Besides, teratogens also regulate genes involved in biological processes that are not restricted to embryogenesis, such as cleavage of cell adhesion proteins as well as lipid metabolism genes induced by SRBEF/SREBP. Although teratogens do not form a homogeneous group and have distinct pharmacological profiles (Figure 5C and Supplementary Figure 5a), these results suggest that transcriptional regulation by teratogens can manifest in changes of several biological processes, potentially mediated by key transcriptional factors. Gene-set enrichment analysis based on *TeraTox* data also revealed molecular insights of teratogenic effects at yet another level of gene expression regulation (Supplementary Figure 7).

In summary, we explain how the *TeraTox* model operates by complementing the machine-learning model with feature importance analysis, pharmacological profiling clustering, sensitivity analysis, and gene-set enrichment analysis.

DISCUSSIONS

This study characterizes the optimization of *TeraTox* and its application in the context of preclinical teratogenicity assessment of drugs. *TeraTox* extends and standardizes the previously published embryoid-body models and fully leverages the predictive potential of these models by adding a toxicological prediction model (Krug *et al.*, 2013; Shinde *et al.*, 2016). It exploits an explainable machine-learning approach to predict teratogenicity potential induced by drug-like molecules.

Only a few pharmaceutical compounds have been recognized as human teratogens based on high-quality data. Therefore, we decided to limit the training data of the *TeraTox* model to a compound set with well-described evidence without losing rigor in the teratogenicity classification ($n = 45$; Table 1). Although alternative compound collections are available, for instance the *ToxCast* data set applied to the *DevTox^{qp}* assay from Stemina (Zurlinden *et al.*, 2020), our compilation was specifically related to pharmaceuticals with teratogenicity profiles supported by either labeling information or strong evidences.

Despite the limited number of reference compounds available, we report that *TeraTox* outperforms 2 QSAR models and performs comparably with state-of-the-art *in-vitro* models. *TeraTox* and *DevTox^{qp}* showed the same accuracy to predict 18 pharmaceutical compounds: the *DevTox^{qp}* assay showed a

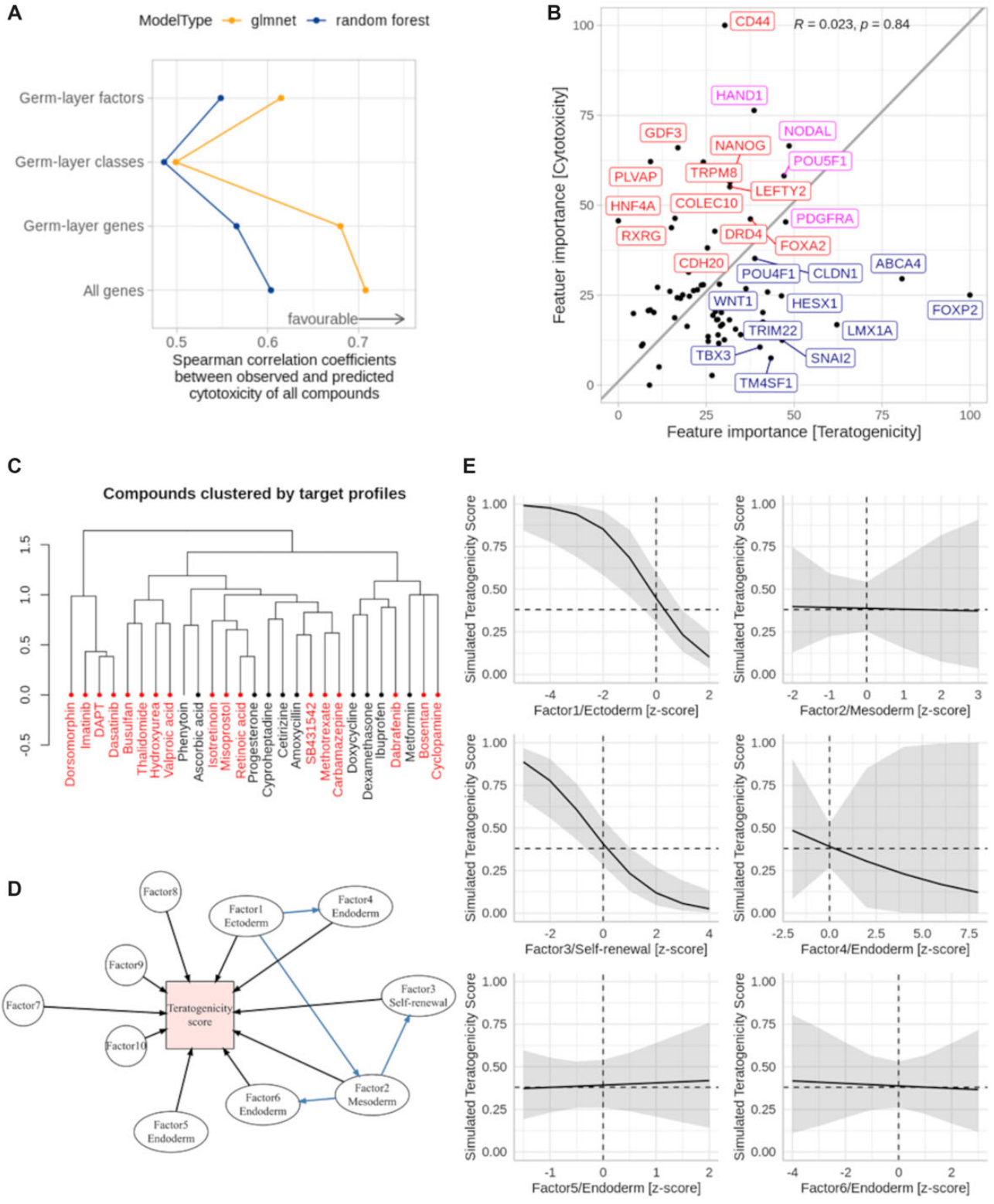

Figure 5. Biological interpretation of the model. A, Prediction of cytotoxicity based on DGE (same workflow as Figure 3A) using all pathway reporter genes as features and an elastic net machine-learning model. Dots: Spearman correlations between observed and predicted cytotoxicity using LOO testing. B, Difference in feature importance of germ-layer genes for prediction of cytotoxicity (upper left) or teratogenicity (lower right) or both (middle). C, Clustering analysis: pharmaceutical target profiles of compounds alone are not sufficient to determine teratogenicity (based on annotated compound target profile, i.e., quantitative affinity to protein targets from ChEMBL database). D, Structure of the DAG to model relationships between teratogenicity score and germ-layer factors with a generalized linear model. Input variables: germ-layer factors and significant interactions between germ-layers identified by Bayesian networks (Factor 1-4 and 6). Model fitting: Supplementary Figure 6. E, Generalized linear model with beta-regression for sensitivity analysis, to test model behavior with tuned input variables. Each panel shows the result of 1 tuning parameter, eg, ectoderm germ-layer factor (top-left) while keeping all other parameters fixed. Lines = average prediction, Areas = 95% confidence intervals of prediction. Input variables are scaled to 0 mean and SD. Abbreviations: DGE, differential gene expression; LOO, leaving-one-out; DAG, directed acyclic graph.

higher specificity and *TeraTox* a higher sensitivity. Combining *TeraTox* and *DevTox*^{qp} led to a higher prediction accuracy, suggesting a synergistic effect of using complementary assays for teratogenicity prediction.

Similarly, we observed comparable performance of *TeraTox* and mEST, with *TeraTox* being more sensitive and mEST more specific. Again, combining both assays led even to an increased accuracy, which would improve overall predictivity.

Due to rich working experience with the mEST assay, we could also offer a first-hand comparison between *TeraTox* and mEST with regard to manual work, cost, and throughput (Supplementary Table 10). We conclude that the overall effort and cost entailed by *TeraTox* is comparable to that of the mEST. However, *TeraTox* offers a much higher throughput thanks to automation and miniaturization, and it generates quantitative gene expression data that can be used to compare new chemical series with existing ones and to further refine the model.

The prediction model presented here is geared towards hazard identification, similar to animal studies, where maximum tolerated doses are to be used for studies testing DART. In the context of an overall risk assessment, one immediate step would be to consider compound potency and to evaluate whether the *in vitro* toxic concentrations would be relevant to human exposure situations. To exemplify this, we highlight here the FP (Supplementary Tables 2 and 5).

Two of them, progesterone and cetirizine, have a minimal transcriptome-altering concentration that is larger than 2 orders of magnitude higher than the human max plasma concentration. A third compound, RO-2, is likely to be similarly "overdosed" *in vitro*. And also, ascorbic acid gives positive signals only at clearly higher concentrations than usually found in human plasma. These considerations could be made more exactly on the basis of physiologically based pharmacokinetic models and considerations of free drug concentrations. However, the principle is exemplified here, and if eg, 4 of the FP would be negative at clinically realistic concentrations, the test specificity would be 76%.

TeraTox offers additional benefits that are not yet available in any existing models. First, *TeraTox* is more sensitive than either the mEST or the *DevTox*^{qp} assay, especially when we consider maximum plasma concentrations (C_{max}) from human data whenever possible or model species otherwise (Supplementary Tables 2 and 3). Detecting human-specific teratogens are critical for drug discovery and development, as illustrated by phthalimide-based series of molecules, which includes thalidomide (Belair *et al.*, 2020; Donovan *et al.*, 2018; Matyskiela *et al.*, 2018; Smith and Mitchell, 2018). Thalidomide was correctly identified as positive by *TeraTox*. Second, *TeraTox* reveals concentration-response teratogenicity and cytotoxicity relationship. This can be integrated with pharmacokinetic and exposure data to better estimate teratogenic risk in the clinic. Third, *TeraTox* generates quantitative gene expression data. Here, we used these data to reveal germ-layer factors, to build a predictive model, and to identify pathways and gene-sets that are regulated by teratogens differently than by nonteratogens. Gene expression data can be also used to explore mechanisms of action and to prioritize drug candidates for preclinical development.

What sets *TeraTox* apart from other models is that it is less of a phenotypic black box but rather an interpretable and explainable model that provides mechanistic insights into gene, pathway, and germ-layer modulations. *TeraTox* informs predictions not only based on statistical data patterns but builds upon biological mechanisms and thus may reflect disturbed functionalities, similar to those leading to teratogenicity *in vivo*. These features put *TeraTox* conceptually in a group of other assays that use phenotypic changes or disturbed functionalities as readouts (Dreser *et al.*, 2020; Hoelting *et al.*, 2016; Meisig *et al.*, 2020; Pallocca *et al.*, 2016). The model consolidates our previous intention to "focus on germ layers" and corroborates recent work exploring gastruloid models that profiles morphological changes of germ-layers for teratogenicity prediction (Jaklin *et al.*, 2020; Moris *et al.*, 2020).

Interpretability and explainability analysis shed light on both the strengths and the limitations of the *TeraTox* model. Most importantly, we could distinguish cytotoxicity from teratogenicity. We explored machine-learning model variants for both teratogenicity and cytotoxicity predictions and made the observation that the best models depend on the target variable. Whereas germ-layer factors and random forest performed best for teratogenicity prediction, the combination of all pathway reporter genes and regularized linear regression with elastic nets showed the best prediction for cytotoxicity (Supplementary Figs. 3a and 3b). There are 2 likely reasons. First, the molecular phenotyping platform contains well-curated genes that reflect cytotoxicity and cell death, which were highlighted in a previous drug screening study using iPS-derived cardiomyocytes (Drawnel *et al.*, 2017). Therefore, we anticipate that these genes are used by linear regression to predict cytotoxicity. Second, teratogenicity is notably complex. It can be induced in many different ways, with different perturbations leading to different down-stream changes that are collectively known as teratogenicity. Therefore, a change in the total output of the germ-layer regulatory network is probably a more robust readout of teratogenicity than individual genes. Random forest, which is an ensemble learning method, is better at detecting such heterogeneous signals than linear regression.

Further studies are warranted to explore several parallel paths for further optimization of the *TeraTox* assay. These can be divided into 4 categories: (1) paths leading to better characterization of EB differentiation, (2) paths leading to testing of larger chemical spaces beyond pharmaceuticals, (3) paths leading to better predictive and explanatory algorithms, and (4) paths leading to better biological models of human embryo development. These lines of research could broaden the applicability domain and increase the robustness of the *TeraTox* assay. To better characterize EBs, multimodal characterizations of the EBs using bulk and single-cell omics, morphological profiling, and time-series experiments could be used. Extension of the assay duration to more than 7 days or using other differentiation protocols may further improve *TeraTox* capacity to model mesoderm and endoderm development.

There are several options to further improve the predictivity and the interpretability of the *TeraTox* model. To better distinguish between nonteratogens and teratogens, we may try to test the compounds with the *TeraTox* assay at lower concentrations (especially for nonteratogens), where the lowest concentration should be predicted to have a TS equal to or close to zero. In this context, it could be feasible to apply the exposure-based validation approach described by Daston *et al.* (2014), based on minimal and maximal concentration-dependent effects of teratogenicity. Another option could be to include the *TeraTox* assay in a test battery to preselect those compounds that show high cytotoxic interference with weak teratogenicity. Multimodel data could be used to identify further relevant features beyond germ-layer genes and factors. As more data are collected, we may also optimize the prediction algorithm, for instance using the nearest-neighbor prediction or other variants.

Finally, the *TeraTox* assay may benefit from a better modeling of human embryo development. We may use alternative morphology-based assays of gastruloids to complement the *TeraTox* readout (Baillie-Benson *et al.*, 2020; Moris *et al.*, 2020). Alternatively, sophisticated microphysiological systems may better mimic the maternal-placenta-embryo axis and hence, may recapitulate true embryo exposure levels and give insights into active drug metabolism although drugs do not need to cross the placental barrier to cause fetal harm (Blundell *et al.*, 2016, 2018; Boos *et al.*, 2021). In the future, they may replace the 3D EBs in *TeraTox*. In the current throughput, though, such systems will probably be more powerful as a secondary assay to spot check a few compounds of particular interest. For this purpose, a continuous integration and modeling of data of human embryogenesis, for instance from omics, imaging, and perturbation studies, is required to guide further optimization of the *TeraTox* assay (Canzler *et al.*, 2020; Mantziou *et al.*, 2021; Yan *et al.*, 2013).

CONCLUSION

We demonstrate that the *TeraTox* is a novel predictive human *in vitro* assay for pharmaceutical teratogenicity prediction that addresses several limitations of alternative assays regarding sensitivity, species-specificity, and interpretability. We believe that its adoption in drug discovery empowers preclinical teratogenicity assessment. Further optimization of the *TeraTox* assay and its routine use in drug-screening processes will lead us towards better preclinical assessment of teratogenicity. Thus, we solicit the community for helping us with further refining and validating *TeraTox* in drug discovery and other contexts.

SUPPLEMENTARY DATA

Supplementary data are available at *Toxicological Sciences* online.

DATA AVAILABILITY

Sequencing files and gene counts are available at the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183534; last access May 3, 2022). Differential gene expression data of commercially available compounds as well as the composition of germ-layer factors are available on the Zenodo platform (https://zenodo.org/record/6143691; last access May 3, 2022).

ACKNOWLEDGMENTS

We thank Kevin Michaelsen, Claudia Bossen, Jean-Christophe Hoflack, Marc Sultan, and Fabian Birzele for their generous support, as well as colleagues of the Bioinformatics and Exploratory Data Analysis (BEDA) team for their input and discussions.

FUNDING

This work has been supported by CEFIC (European Chemical Industry Council), the BMBF (Bundesministerium für Bildung und Forschung), EFSA (European Food Safety Authority), and the DK-EPA (Danish Environmental Protection Agency) (MST-667-00205). It has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No. 681002 (EU-ToxRisk), No. 964537 (RISK-HUNT3R), No. 964518 (ToxFree), and No. 825759 (ENDpoiNTs) and from Horizon Europe.

DECLARATION OF CONFLICTING INTERESTS

M.J. is currently employed at Weleda AG and some authors (J.D.Z., S.P.L., N.S., P.B., E.K., N.C., and S.K.) are employees of F. Hoffmann-La Roche Ltd. All authors declare that they have no conflicts of interest.

REFERENCES

Adams, S. S., Bough, R. G., Cliffe, E. E., Lessel, B., and Mills, R. F. N. (1969). Absorption, distribution and toxicity of ibuprofen. *Toxicol. Appl. Pharmacol.* **15**, 310–330.

Adler, S., Pellizzer, C., Hareng, L., Hartung, T., and Bremer, S. (2008). First steps in establishing a developmental toxicity test method based on human embryonic stem cells. *Toxicol. In Vitro* **22**, 200–211.

Augustyniak, J., Bertero, A., Coccini, T., Baderna, D., Buzanska, L., and Caloni, F. (2019). Organoids are promising tools for species-specific in vitro toxicological studies. *J. Appl. Toxicol.* **39**, 1610–1622.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., and Zhang, J. D. (2020). An introduction to machine learning. *Clin. Pharmacol. Ther.* **107**, 871–885.

Baillie-Benson, P., Moris, N., and Martinez Arias, A. (2020). Pluripotent stem cell models of early mammalian development. *Curr. Opin. Cell Biol.* **66**, 89–96.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., *et al.*, (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115.

Barrow, P. (2016). Revision of the ICH guideline on detection of toxicity to reproduction for medicinal products: SWOT analysis. *Reprod. Toxicol.* **64**, 57–63.

Beck, F., Erler, T., Russell, A., and James, R. (1995). Expression of cdx-2 in the mouse embryo and placenta: Possible role in patterning of the extra-embryonic membranes. *Dev. Dyn.* **204**, 219–227.

Belair, D. G., Lu, G., Waller, L. E., Gustin, J. A., Collins, N. D., and Kolaja, K. L. (2020). Thalidomide inhibits human iPSC mesendoderm differentiation by modulating CRBN-dependent degradation of SALL4. *Sci. Rep.* **10**, 2864.

Blundell, C., Tess, E. R., Schanzer, A. S., Coutifaris, C., Su, E. J., Parry, S., and Huh, D. (2016). A microphysiological model of the human placental barrier. *Lab. Chip.* **16**, 3065–3073.

Blundell, C., Yi, Y. S., Ma, L., Tess, E. R., Farrell, M. J., Georgescu, A., Aleksunes, L. M., and Huh, D. (2018). Placental drug transport-on-a-chip: A microengineered in vitro model of transporter-mediated drug efflux in the human placental barrier. *Adv. Healthc. Mater.* **7**, 1700786.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., Ziller, M., Croft, G. F., Amoroso, M. W., Oakley, D. H., *et al.*, (2011). Reference maps of human ES and IPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452.

Boos, J. A., Misun, P. M., Brunoldi, G., Furer, L. A., Aengenheister, L., Modena, M., Rousset, N., Buerki-Thurnherr, T., and Hierlemann, A. (2021). Microfluidic co-culture platform to recapitulate the maternal-placental-embryonic axis. *Adv. Biol. (Weinh.)* **5**, e2100609.

Brooks, M., Kristensen, K., van Benthem, K., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H., Mächler, M., and Bolker, B. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* **9**, 378–400.

Burridge, P. W., Thompson, S., Millrod, M. A., Weinberg, S., Yuan, X., Peters, A., Mahairaki, V., Koliatsos, V. E., Tung, L., and Zambidis, E. T. (2011). A universal system for highly efficient cardiac differentiation of human induced pluripotent stem cells that eliminates interline variability. *PLoS One* **6**, e18293.

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., and Hackermüller, J. (2020). Prospects and challenges of multi-omics data integration in toxicology. *Arch. Toxicol.* **94**, 371–388.

Cassano, A., Manganaro, A., Martin, T., Young, D., Piclin, N., Pintore, M., Bigoni, D., and Benfenati, E. (2010). Caesar models for developmental toxicity. *Chem. Cent. J.* **4** (Suppl 1), S4.

Chen, J. K., Taipale, J., Cooper, M. K., and Beachy, P. A. (2002). Inhibition of hedgehog signaling by direct binding of cyclopamine to smoothened. *Genes Dev.* **16**, 2743–2748.

Cusack, B. J., Parsons, T. E., Weinberg, S. M., Vieira, A. R., and Szabo-Rogers, H. L. (2017). Growth factor signaling alters the morphology of the zebrafish ethmoid plate. *J. Anat.* **230**, 701–709.

Daniel, S., Doron, M., Fishman, B., Koren, G., Lunenfeld, E., and Levy, A. (2019). The safety of amoxicillin and clavulanic acid use during the first trimester of pregnancy. *Br. J. Clin. Pharmacol.* **85**, 2856–2863.

Dashe, J. S., and Gilstrap, L. C. 3rd (1997). Antibiotic use in pregnancy. *Obstetr. Gynecol. Clin. N. Am.* **24**, 617–629.

Daston, G. P., Beyer, B. K., Carney, E. W., Chapin, R. E., Friedman, J. M., Piersma, A. H., Rogers, J. M., and Scialli, A. R. (2014). Exposure-based validation list for developmental toxicity screening assays. *Birth Defects Res. B Dev. Reprod. Toxicol.* **101**, 423–428.

Donovan, K. A., An, J., Nowak, R. P., Yuan, J. C., Fink, E. C., Berry, B. C., Ebert, B. L., and Fischer, E. S. (2018). Thalidomide promotes degradation of sall4, a transcription factor implicated in Duane Radial Ray syndrome. *eLife* **7**, e38430.

Drawnel, F. M., Zhang, J. D., Küng, E., Aoyama, N., Benmansour, F.A., Del Rosario, A., Jensen Zoffmann, S., Delobel, F., Prummer, M., Weibel, F., *et al.* (2017). Molecular phenotyping combines molecular information, biological relevance, and patient data to improve productivity of early drug discovery. *Cell Chem. Biol.* **24**, 624–634.e623.

Dreser, N., Madjar, K., Holzer, A. K., Kapitza, M., Scholz, C., Kranaster, P., Gutbier, S., Klima, S., Kolb, D., Dietz, C., *et al.* (2020). Development of a neural rosette formation assay (RoFA) to identify neurodevelopmental toxicants and to characterize their transcriptome disturbances. *Arch. Toxicol.* **94**, 151–171.

Etwel, F., Djokanovic, N., Moretti, M. E., Boskovic, R., Martinovic, J., and Koren, G. (2014). The fetal safety of cetirizine: An observational cohort study and meta-analysis. *J. Obstetrics Gynaecol.* **34**, 392–399.

Evans, T. J. 2007. Chapter 14 - reproductive toxicity and endocrine disruption. In *Veterinary Toxicology* (R. C. Gupta, Ed.), pp. 206–244. Academic Press, Oxford.

Genschow, E., Spielmann, H., Scholz, G., Pohl, I., Seiler, A., Clemann, N., Bremer, S., and Becker, K. (2004). Validation of the embryonic stem cell test in the international ECVAM validation study on three in vitro embryotoxicity tests. *Altern. Lab. Anim.* **32**, 209–244.

Genschow, E., Scholz, G., Brown, N., Piersma, A., Brady, M., Clemann, N., Huuskonen, H., Paillard, F., Bremer, S., Becker, K., *et al.* (2000). Development of prediction models for three in vitro embryotoxicity tests in an ECVAM validation study. *In Vitro Mol. Toxicol.* **13**, 51–66.

Hoelting, L., Klima, S., Karreman, C., Grinberg, M., Meisig, J., Henry, M., Rotshteyn, T., Rahnenführer, J., Blüthgen, N., Sachinidis, A., *et al.* (2016). Stem cell-derived immature human dorsal root ganglia neurons to identify peripheral neurotoxicants. *Stem Cells Transl. Med.* **5**, 476–487.

ICH S5 (R3): International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (2020) Detection of Toxicity to Reproduction for Human Pharmaceuticals S5(R3). https://www.ema.europa.eu/en/documents/scientific-guideline/ich-s5-r3-guideline-reproductive-toxicology-detection-toxicity-reproduction-human-pharmaceuticals_en.pdf (accessed May 3, 2022).

Jaklin, M., Zhang, J. D., Barrow, P., Ebeling, M., Clemann, N., Leist, M., and Kustermann, S. (2020). Focus on germ-layer markers: A human stem cell-based model for in vitro teratogenicity testing. *Reprod. Toxicol.* **98**, 286–298.

Kameoka, S., Babiarz, J., Kolaja, K., and Chiao, E. (2014). A high-throughput screen for teratogens using human pluripotent stem cells. *Toxicol. Sci.* **137**, 76–90.

Krug, A. K., Kolde, R., Gaspar, J. A., Rempel, E., Balmer, N. V., Meganathan, K., Vojnits, K., Baquie, M., Waldmann, T., Ensenat-Waser, R., *et al.* (2013). Human embryonic stem cell-derived test systems for developmental neurotoxicity: A transcriptomics approach. *Arch. Toxicol.* **87**, 123–143.

Lipinski, R. J., Hutson, P. R., Hannam, P. W., Nydza, R. J., Washington, I. M., Moore, R. W., Girdaukas, G. G., Peterson, R. E., and Bushman, W. (2008). Dose- and route-dependent teratogenicity, toxicity, and pharmacokinetic profiles of the hedgehog signaling antagonist cyclopamine in the mouse. *Toxicol. Sci.* **104**, 189–197.

Ljosa, V., Caie, P. D., ter Horst, R., Sokolnicki, K. L., Jenkins, E. L., Daya, S., Roberts, M. E., Jones, T. R., Singh, S., Genovesio, A., *et al.* (2013). Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* **18**, 1321–1329.

Lüdecke, D. (2018). Ggeffects: tidy data frames of marginal effects from regression models. *J. Open Source Softw.* **3**, 772.

Mantziou, V., Baillie-Benson, P., Jaklin, M., Kustermann, S., Arias, A. M., and Moris, N. (2021). *In vitro* teratogenicity testing using a 3d, embryo-like gastruloid system. bioRxiv. 2021.2003.2030.437698.

Marzo, M., Kulkarni, S., Manganaro, A., Roncaglioni, A., Wu, S., Barton-Maclaren, T. S., Lester, C., and Benfenati, E. (2016). Integrating in silico models to enhance predictivity for developmental toxicity. *Toxicology* **370**, 127–137.

Matyskiela, M. E., Couto, S., Zheng, X., Lu, G., Hui, J., Stamp, K., Drew, C., Ren, Y., Wang, M., Carpenter, A., *et al.* (2018). Sall4 mediates teratogenicity as a thalidomide-dependent cereblon substrate. *Nat. Chem. Biol.* **14**, 981–987.

Meisig, J., Dreser, N., Kapitza, M., Henry, M., Rotshteyn, T., Rahnenführer, J., Hengstler, J. G., Sachinidis, A., Waldmann, T., Leist, M., *et al.* (2020). Kinetic modeling of stem cell transcriptome dynamics to identify regulatory modules of normal and disturbed neuroectodermal differentiation. *Nucleic Acids Res.* **48**, 12577–12592.

Moris, N., Anlas, K., van den Brink, S. C., Alemany, A., Schröder, J., Ghimire, S., Balayo, T., van Oudenaarden, A.M., and Arias, A. (2020). An in vitro model of early anteroposterior organization during human development. *Nature* **582**, 410–415.

Muanda, F. T., Sheehy, O., and Bérard, A. (2017). Use of antibiotics during pregnancy and the risk of major congenital malformations: A population based cohort study. *Br. J. Clin. Pharmacol.* **83**, 2557–2571.

Pallocca, G., Grinberg, M., Henry, M., Frickey, T., Hengstler, J. G., Waldmann, T., Sachinidis, A., Rahnenführer, J., and Leist, M. (2016). Identification of transcriptome signatures and biomarkers specific for potential developmental toxicants inhibiting human neural crest cell migration. *Arch. Toxicol.* **90**, 159–180.

Palmer, J. A., Smith, A. M., Egnash, L. A., Colwell, M. R., Donley, E. L. R., Kirchner, F. R., and Burrier, R. E. (2017). A human induced pluripotent stem cell-based in vitro assay predicts developmental toxicity through a retinoic acid receptor-mediated pathway for a series of related retinoid analogues. *Reprod. Toxicol.* **73**, 350–361.

Palmer, J. A., Smith, A. M., Egnash, L. A., Conard, K. R., West, P. R., Burrier, R. E., Donley, E. L., and Kirchner, F. R. (2013). Establishment and assessment of a new human embryonic stem cell-based biomarker assay for developmental toxicity screening. *Birth Defects Res. B Dev. Reprod. Toxicol.* **98**, 343–363.

Quintanilla, R. H., Jr, Asprer, J. S. T., Vaz, C., Tanavde, V., and Lakshmipathy, U. (2014). Cd44 is a negative cell surface marker for pluripotent stem cell identification during human fibroblast reprogramming. *PLoS One* **9**, e85419.

Rempel, E., Hoelting, L., Waldmann, T., Balmer, N. V., Schildknecht, S., Grinberg, M., Das Gaspar, J. A., Shinde, V., Stober, R., Marchan, R., *et al.* (2015). A transcriptome-based classifier to identify developmental toxicants by stem cell testing: Design, validation and optimization for histone deacetylase inhibitors. *Arch. Toxicol.* **89**, 1599–1618.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.

Rumbold, A., Ota, E., Nagata, C., Shahrook, S., and Crowther, C. A.; Cochrane Pregnancy and Childbirth Group. (2015). Vitamin c supplementation in pregnancy. *Cochrane Database Syst. Rev.* **2016**, CD004073.

Sakata, T., and Chen, J. K. (2011). Chemical 'jekyll and hyde's: Small-molecule inhibitors of developmental signaling pathways. *Chem. Soc. Rev.* **40**, 4318–4331.

Scholz, G., Genschow, E., Pohl, I., Bremer, S., Paparella, M., Raabe, H., Southee, J., and Spielmann, H. (1999a). Prevalidation of the embryonic stem cell test (EST)-a new in vitro embryotoxicity test. *Toxicol. In Vitro* **13**, 675–681.

Scholz, G., Pohl, I., Genschow, E., Klemm, M., and Spielmann, H. (1999b). Embryotoxicity screening using embryonic stem cells in vitro: Correlation to in vivo teratogenicity. *Cells Tissues Organs* **165**, 203–211.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.* **35**, 1–22.

Shinde, V., Hoelting, L., Srinivasan, S. P., Meisig, J., Meganathan, K., Jagtap, S., Grinberg, M., Liebing, J., Bluethgen, N., Rahnenfuhrer, J., *et al.* (2017). Definition of transcriptome-based indices for quantitative characterization of chemically disturbed stem cell development: Introduction of the stop-Toxukn and stop-Toxukk tests. *Arch. Toxicol.* **91**, 839–864.

Shinde, V., Klima, S., Sureshkumar, P. S., Meganathan, K., Jagtap, S., Rempel, E., Rahnenfuhrer, J., Hengstler, J. G., Waldmann, T., Hescheler, J., *et al.* (2015). Human pluripotent stem cell based developmental toxicity assays for chemical safety screening and systems biology data generation. *J. Vis. Exp.* **100**, e52333.

Shinde, V., Perumal Srinivasan, S., Henry, M., Rotshteyn, T., Hescheler, J., Rahnenfuhrer, J., Grinberg, M., Meisig, J., Bluthgen, N., Waldmann, T., *et al.* (2016). Comparison of a teratogenic transcriptome-based predictive test based on human embryonic versus inducible pluripotent stem cells. *Stem Cell Res. Ther.* **7**, 190.

Smith, R. L., and Mitchell, S. C., (2018). Thalidomide-type teratogenicity: structure activity relationships for congeners. *Toxicology Research*, **7**, 1036–1047.

Tsankov, A. M., Akopian, V., Pop, R., Chetty, S., Gifford, C. A., Daheron, L., Tsankova, N. M., and Meissner, A. (2015a). A qPCR scorecard quantifies the differentiation potential of human pluripotent stem cells. *Nat. Biotechnol.* **33**, 1182–1192.

Tsankov, A. M., Gu, H., Akopian, V., Ziller, M. J., Donaghey, J., Amit, I., Gnirke, A., and Meissner, A. (2015b). Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344–349.

Waldmann, T., Grinberg, M., König, A., Rempel, E., Schildknecht, S., Henry, M., Holzer, A.-K., Dreser, N., Shinde, V., Sachinidis, A., *et al.* (2017). Stem cell transcriptome responses and corresponding biomarkers that indicate the transition from adaptive responses to cytotoxicity. *Chem. Res. Toxicol.* **30**, 905–922.

Waldmann, T., Rempel, E., Balmer, N. V., Konig, A., Kolde, R., Gaspar, J. A., Henry, M., Hescheler, J., Sachinidis, A., Rahnenfuhrer, J., et al (2014). Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. *Chem. Res. Toxicol.* **27**, 408–420.

Wang, X., Moon, J., Dodge, M. E., Pan, X., Zhang, L., Hanson, J. M., Tuladhar, R., Ma, Z., Shi, H., Williams, N. S., *et al.* (2013). The development of highly potent inhibitors for porcupine. *J. Med. Chem.* **56**, 2700–2704.

Whitlow, S., Bürgin, H., and Clemann, N. (2007). The embryonic stem cell test for the early selection of pharmaceutical compounds. *ALTEX* **24**, 3–7.

Worley, K. E., Rico-Varela, J., Ho, D., and Wan, L. Q. (2018). Teratogen screening with human pluripotent stem cells. *Integr. Biol. (Camb.)* **10**, 491–501.

Wu, S., Fisher, J., Naciff, J., Laufersweiler, M., Lester, C., Daston, G., and Blackburn, K. (2013). Framework for identifying chemicals with structural features associated with the potential to act as developmental or reproductive toxicants. *Chem. Res. Toxicol.* **26**, 1840–1861.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., *et al.* (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139.

Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G.-W., Tao, C. Y., Tallarico, J. A., Labow, M., Jenkins, J. L., Mitchison, T. J., *et al.* (2008). Integrating high-content screening and

ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.* **4**, 59–68.

Zhang, J. D., Hatje, K., Sturm, G., Broger, C., Ebeling, M., Burtin, M., Terzi, F., Pomposiello, S. I., and Badi, L. (2017). Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genomics* **18**, 277.

Zhang, J. D., Küng, E., Boess, F., Certa, U., and Ebeling, M. (2015). Pathway reporter genes define molecular phenotypes of human cells. *BMC Genomics* **16**, 342.

Zhang, J. D., Schindler, T., Küng, E., Ebeling, M., and Certa, U. (2014). Highly sensitive amplicon-based transcript quantification by semiconductor sequencing. *BMC Genomics* **15**, 565.

Zurlinden, T. J., Saili, K. S., Rush, N., Kothiya, P., Judson, R. S., Houck, K. A., Hunter, E. S., Baker, N. C., Palmer, J. A., Thomas, R. S., *et al.* (2020). Profiling the toxcast library with a pluripotent human (h9) stem cell line-based biomarker assay for developmental toxicity. *Toxicol. Sci.* **174**, 189–209.