Optimization of the *TeraTox* assay for preclinical teratogenicity assessment

*Jaklin Manuela**[,1,2,4], *Zhang Jitao David**[,1], *Schäfer Nicole*[1], *Clemann Nicole*[1], *Barrow Paul*[1], *Küng Erich*[1], *Sach-Peltason Lisa*[1], *McGinnis Claudia*[3], *Leist Marcel*[2], *Kustermann Stefan*[1]

[1] Pharma Research and Early Development, Roche Innovation Centre Basel,

F. Hoffmann - La Roche, Switzerland

[2] Department for In Vitro Toxicology and Biomedicine Inaugurated by the Doerenkamp-Zbinden Foundation, University of Konstanz, Germany

[3] University of Dundee, Drug Discovery Unit, Dundee, Scotland, UK

[4] Present Address: Weleda AG, Arlesheim, Switzerland

*these authors contributed equally

To whom correspondence should be addressed: mail.manuela.jaklin@gmail.com (MJ), jitao_david.zhang@roche.com (JDZ).

Abstract

Teratogenicity poses severe threats to patient safety. Stem-cell-based *in vitro* systems are promising tools to predict human teratogenicity. However, current *in vitro* assays are limited because they either capture effects on a certain germ layer, or focus on a subset of predictive markers. Here we report the characterization and critical assessment of *TeraTox*, a newly developed multi-lineage differentiation assay using 3D human induced pluripotent stem cells. *TeraTox* probes stem-cell derived embryoid bodies with two endpoints, one quantifying cytotoxicity and the other inferring the teratogenicity potential with gene expression as a molecular phenotypic readout. To derive teratogenicity potentials from gene expression profiles, we applied both unsupervised machine-learning tools including factor analysis and supervised tools including classification and regression. To identify the best predictive model for the teratogenicity potential that is explainable, we systematically tested 64 machine-learning model architectures and identified the optimal model, which uses expression of 77 representative germ-layer genes, summarized by 10 latent germ-layer factors, as input for random-forest regression. We combined measured cytotoxicity and inferred teratogenicity potential to predict concentration-dependent teratogenicity profiles of 33 approved pharmaceuticals and 12 proprietary drug candidates with known *in vivo* data. Compared with the mouse embryonic stem cell test, which has been in routine use for more than a decade, the *TeraTox* assay shows higher sensitivity, particularly towards teratogens impairing ectodermal development or stem-cell renewal, and a more balanced prediction performance. We envision that further refinement and development of *TeraTox* has the potential to reduce and replace animal research in drug discovery and to improve preclinical assessment of teratogenicity.

Keywords: *teratogenicity, molecular phenotyping, mEST, TeraTox, stem cells, embryoid bodies, machine learning*

1. Introduction

Teratogenicity, the ability of a chemical to cause defects in a developing fetus, has gained wide and continuous attention since the thalidomide tragedy in the 1960s (1). To assess the teratogenic potential of drug candidates, pharmaceutical companies must perform embryo-fetal-development studies (EFD studies hereafter) in at least one rodent and one non-rodent species (2,3). There is an urgent need to develop alternative, humanized *in vitro* assays for early assessment of teratogenicity, because they can potentially better mimic human physiology, reduce animal use in drug discovery, and lower the attrition rate of drug development by filtering out potential teratogens early (3-5).

The current industrial standard *in vitro* assay for teratogenicity assessment is the mouse embryonic stem cell test (mEST). It measures both the differentiation of embryoid bodies (EB) derived from D3 mouse embryonic stem cells (mESC) by quantifying beating cardiac tissue, and the cytotoxicity in both mouse D3 ESCs as well as mouse 3T3 fibroblasts (6-8).

The mEST assay offers several advantages compared with other assays, including the zebrafish model (9,10) and other stem-cell-based *in vitro* models (11-18). It uses two well-characterized, stable cell lines as the biological model that recapitulates early embryogenesis and no animal experiments are required. The cells are easy to acquire and handle. The protocol is well established and the assay is widely adopted. Importantly, the assay is validated by the European Centre for the Validation of Alternative Methods (7,8,19-21), and therefore trusted by many laboratories.

However, the mEST assay has both conceptual and practical limitations as a predictive model of human teratogenicity. First, it uses murine cells, which fail to recapitulate human teratogenicity for some chemical classes, for instance phthalimide-based molecules including thalidomide (22). Second, because the stem cells are differentiated into cardiomyocytes, the assay preferentially quantifies impairment of mesodermal germ-layer development. Third, the EB differentiation is a lengthy process of ten days and the manual counting of beating cardiomyocytes is both time-consuming and error-prone, which limits the throughput of the assay. Finally, and critically, the predictive algorithm relies on ID_{50}, the concentration at which half of the maximal inhibition of differentiation is achieved. For strong cytotoxic compounds, it is common that IC_{50}, the concentration at which half of the maximal cytotoxicity is observed, coincides closely with ID_{50}, which causes false-negative predictions. Since the assay is used to pre-select developmental compounds prior to regulatory EFD studies, misclassifications necessitate unnecessary animal use in EFD studies and, in case the teratogenicity is specific to humans, pose severe threats for patient safety.

Given the limitations of the mEST assay, we developed a new, humanized *in vitro* teratogenicity assay. The new assay, which we call *TeraTox*, uses ethically non-restricted human induced pluripotent stem cells (hiPSC). The cells form three-dimensional embryoid bodies (EBs) and differentiate spontaneously into all three germ layers – ectoderm, mesoderm, and endoderm – with expression of representative developmental markers of each layer. We previously documented the technical details of the assay and demonstrated its feasibility with four teratogens and four non-teratogens (23). However, a systematic assessment of its performance using a larger compound set has not been conducted yet and the prediction algorithm is missing.

To fill these gaps, here we describe the optimization and critical assessment of the *TeraTox* assay and the setup of a predictive model for human teratogenicity evaluation. We compiled a panel of 45 drug-like molecules with known teratogenicity profiles and tested them in six-point concentration response, generating the largest published dataset so far in a single study about *in vitro* modelling of teratogenicity with reference to clinical/ animal *in vivo* data. Because both the cell amount and the workload required by digital PCR would be prohibitive, we adapted *Molecular Phenotyping,* a technology based on amplicon-based RNA sequencing, to quantify expression of germ-layer genes. Using gene expression data as input, we built machine-learning models with varying architectures and identified the best-performing model using factor analysis and random-forest regression. Using a leave-one-out training-testing strategy, we classified the 45 compounds as either teratogenic or non-teratogenic, thereby considering both concentration-dependent cytotoxicity and teratogenicity potential. We found that *TeraTox* features a lower specificity but outperforms mEST with regard to sensitivity and balanced prediction considering precision and sensitivity. Finally, we augmented the model with biological and pharmacological interpretations as well as simulation studies that explain how it works. In summary, our assessment highlights both the advantage of *TeraTox* over the standard mEST assay for preclinical teratogenicity assessment and directions of its future development.

2. Material and Methods

2.1. Human iPSC derived *TeraTox* Assay

The *TeraTox* assay is built upon commercially available human induced pluripotent stem cells (hiPSC, Gibco, A18945) with indistinguishable gene expression profiles compared with embryonic stem cells (16,24). The cells form 3D EBs and undergo multi-lineage differentiation into all three germ layers (23). Prior to the assay, the hiPSC were tested with the TaqMan ScoreCard assay (Thermo Fisher) to confirm sufficient levels of pluripotency (25). The EBs were spontaneously differentiated and treated with several reference substances over a time course of seven 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 day 0, day 3 and day 5 at six concentrations, together with EB medium and 0.25% DMSO solvent controls as the negative reference. Cell viability was determined on day 7 by measuring ATP release in supernatants with the CellTiter-Glo 3D assay (Promega, G9681) to pre-specify appropriate testing ranges. All cell culture media and reagents were obtained from Gibco (Thermo Fisher) unless otherwise specified. The overall cell culture and cytotoxicity protocol was previously described in detail by Jaklin *et al.*, 2020 (23).

Targeted gene expression profiling was performed with the molecular phenotyping platform that we developed previously (26-28). In total, 1,055 samples of differentiated EBs were lysed after 7 days of differentiation in 350 µl MagNA Pure LC RNA Buffer (Roche Diagnostics) and purified by using the 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 the mosquito automatic pipettor (SPT Labtech) in miniaturization. For the purifications before and after final library amplification, solid phase reversible immobilization magnetic bead purification (Clean NGS, LABGENE Scientific SA) was performed on the 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 Biomek FXP workstation. The libraries were sequenced on the NovaSeq 6000 Instrument (Illumina) with the sequencing-by-synthesis technology. All the 75 cycles ended up with a minimum of 2 Mio sequencing reads per sample for analysis. We used molecular phenotyping with 1,215 detectable pathway reporter genes including a subset of 87 early developmental markers (germ-layer genes, Suppl. Tab. S3) and genes representative of toxicological pathways to identify differentially expressed genes induced by the compounds at pre-specified concentration levels (25,29).

2.2. Mouse Embryonic Stem Cell Test

The protocol of the mEST was adapted from the original publication from Genschow *et al*., 2004 (7) into an industry compliant format (8). We used the pluripotent mouse embryonic stem cell line ES-D3 (ATCC, CRL-1934) and the somatic mouse 3T3 fibroblast line (Balb/c 3T3 cell clone A31 from ATCC, CCL-163). Most manual steps of the assay, such as cell seeding, dilution and addition of compounds, centrifugation and incubation of the EBs, are standardized and automated to gain reproducible data (30). The only non-automated assay procedures were cell maintenance and the manual count of beating cardiomyocytes.

The mEST assay is performed in two steps. First, the MTT cytotoxicity assay (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) is conducted with both differentiated 3T3 fibroblasts and pluripotent D3 ESCs in monolayer cultures. Second, EBs derived from D3 ESCs are differentiated into cardiomyocytes over a total time course of 10 days, with compound treatment in six different concentrations on day 4 and day 7. The endpoints measured are the concentration at which 50% inhibition of growth of 3T3 (IC_{50} 3T3) and D3 cells (IC_{50} D3) is achieved, and the concentration at which 50% inhibition of differentiation into cardiomyocytes (ID_{50} D3) is achieved, compared to DMSO solvent controls, respectively (Suppl. Fig. S1a).

A modified discriminant function analysis was used to classify the test chemicals into two groups based on the calculated predictive score (PS) for low potential of teratogenicity (negative, PS <0.6) and high potential of teratogenicity (positive, PS ≥0.6).

A possible prediction result is ‘borderline’, if calculated predictive scores are below the cut-off of 0.6 but above 0.5. Inconclusive results are also possible, for example, if solubility limits the concentration ranges tested to an extent that no IC_{50} or ID_{50} values can be reliably determined for one or more concentration response curves (Suppl. Fig. S1b).

2.3. Assessing characteristics of differentiated hiPSC with BioQC

We applied the *BioQC* software that we developed previously to characterize the identity of the differentiated samples across all treated compound concentrations (including vehicle controls) at the endpoint on day 7 (31). 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 collected from public compendia (32,33). The BioQC performs Wilcoxon-Mann-Whitney tests comparing expression of genes in a set, for instance genes preferentially expressed in one 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.

2.4. Analysis and modelling of the *TeraTox* data

We performed differential gene expression analysis comparing compound-treated samples with DMSO controls using the generalized linear model implemented in the *edgeR* package in R/Bioconductor (34). 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, given by the sign of log2 fold-change (logFC). The vectors of *z*-scores of all genes (N=1,215) were used as raw features for machine-learning models, based on which further feature selection and engineering work was performed.

We also tested the possibility of using the effect size, logFC, as features. However, we found that using *z*-scores as features delivered better generalizability between training and testing datasets. Therefore, we report the performance of models using *z*-scores unless otherwise specified.

Besides the raw feature set of *z*-scores of all genes, we used three 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 (Suppl. Tab. S3) (23,25,29,35). Second, we used the germ-layer associations reported by Tsankov *et al.* to derive a reduced feature set defined by five germ-layer classes, including both germ layers (ectoderm, endoderm, mesoderm, mesendoderm) and pluripotency, by taking the median *z*-scores of germ-layer genes associated with each germ-layer class (25). 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 the teratogenicity potential in two 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 define a compound-specific Teratogenicity Score (TS hereafter). For non-teratogens, the TS is defined as 0 independent of the tested concentration. For teratogens, the TS is defined as the 0-1-bounded cosine similarity between the differential expression profile induced by a particular concentration of a certain compound and the differential expression profile induced by the highest non-cytotoxic concentration of the same compound.

The non-cytotoxic concentration was determined by the highest concentration that we tested which is associated with an average variability equal or larger than 80%.

The models were trained and validated using the Leave-One-Out (LOO) scheme. We iterated over all compounds, leaving one compound out at a time and using the remaining compounds to build machine-learning models. Then we compared teratogenicity scores predicted by the models with the observation of the left-out compound with the Spearman correlation coefficient.

As an alternative to LOO, we also tested repeated 80%/20% splitting of data into training sets and test sets. However, it cannot be used to predict teratogenic scores for any particular compound without using its data in both training and testing sets. Therefore, we report only results derived from the LOO scheme unless otherwise stated.

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

Besides predicting teratogenicity scores, we also exhaustively 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 differ from that for teratogenicity scores (further discussed in results).

All data analysis was performed with R (version 4.0.1) or Python (version 3.8.1) unless otherwise specified.

2.5. Test chemicals for validation

In total we tested 27 positive and 18 negative reference substances in six-point concentrations in the mEST and the human *TeraTox* assay (Tab. 1). This compound panel consisted of both commercial and developmental pharmaceuticals with known teratogenicity profiles available from either human evidence-based information, where unambiguous warnings have been found and use during pregnancy is explicitly contraindicated by the FDA, or from *in vivo* EFD studies in rats and/ or rabbits (3,36-47). Compounds without existing human or *in vivo* animal data were classified as teratogens based on known teratogenic risks associated with their mode of action (18,48-55). Some compounds have been taken by cohorts of pregnant women and did not lead to any observed increase in the frequency of malformations during early pregnancy. We considered these drugs as non-teratogenic in humans, at least in the physiologically relevant concentrations of exposure (56-72) (Suppl. Tab. S1).

The commercial compounds were obtained from Merck, Germany. We also included 12 developmental small molecules RO-1 to RO-12 provided by F. Hoffmann – La Roche, Switzerland (compound structures are not disclosed due to confidentiality and intellectual property). Those compounds have unknown human teratogenicity profiles, but *in vivo* data are available from EFD studies performed either in rats, or in rabbits, or in both (Suppl. Tab. S2).

We assigned RO-1, RO-3, RO-8, RO-9 and RO-10 due to the outcome of *in vivo* studies as positive teratogens, and RO-2, RO-4, RO-5, RO-6, RO-7, RO-11, RO-12 as non-teratogens (73). All compounds were serially diluted in DMSO (0.25%) from a stock solution to six test concentrations.

We used the following metrics to compare the performance of the *TeraTox assay* and that of the *mEST* assay. We calculated assay sensitivity as the proportion of correctly predicted teratogens. Assay specificity was calculated as the proportion of correctly predicted non-teratogens. Overall accuracy was taken as the proportion of all correct predictions, and F_1 scores are calculated as the harmonic mean of precision and recall. When we denote True Positive, True Negative, False Positive, and False Negative with *TP*, *TN*, *FP*, and *FN*, respectively, the metrics of performance are defined in Equations 1-5.

$$Recall\ (sensitivity) = \frac{TP}{TP+FN} \qquad \text{(Eq. 1)}$$

$$Specificity = \frac{TN}{TN+FP} \qquad \text{(Eq. 2)}$$

$$Precision = \frac{TP}{TP+FP} \qquad \text{(Eq. 3)}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad \text{(Eq. 4)}$$

$$F_1 = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} \qquad \text{(Eq. 5)}$$

To identify the threshold of TS that maximizes the performance (F_1 score) of the *TeraTox Score* model, we used 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 defined in Equations (1)-(5).

Table 1: Reference compounds used for assay validation, with human teratogenicity labels and test concentration in both human and mouse model. Teratogenicity classification was based on FDA classification (Suppl. Tab. S1) or *in vivo* EFD data (indicated with asterisks*, Suppl. Tab. S2).

Reference Compound	***Teratogenicity Classification***	***Test Concentrations (human model) [µM]***	***Test Concentrations (mouse model) [µM]***
Acitretin	**Positive**	0.08 – 2.5	0.004 – 100
Amoxicillin	Negative	6.25 – 200	39 – 2500
Artesunate	**Positive**	0.13 – 4	0.016 – 100
Ascorbic Acid	Negative	28 – 900	0.035 – 2000
Bosentan	**Positive**	4.7 – 150	7.8 – 500
Busulfan	**Positive**	0.13 – 4	0.6 – 500
Carbamazepine	**Positive**	9.4 – 300	11.7 – 750
Cetirizine	Negative	19 – 600	11.7 – 750
Cyclopamine	**Positive**	0.6 – 20	0.07 – 50
Cyproheptadine	Negative	0.9 – 30	0.3 – 250
Dabrafenib	**Positive**	0.06 – 2	0.1 – 100
DAPT	**Positive**	0.09 – 3	0.1 – 500
Dasatinib	**Positive**	0.6 – 20	0.3 – 20
Dexamethasone	Negative	9.4 – 300	0.3 – 1000
Dorsomorphin	**Positive**	0.4 – 14	0.04 – 50
Doxycycline	Negative	0.6 – 20	1.6 – 1500
5-Fluorouracil	**Positive**	0.08 – 0.25	0.02 – 20
Hydroxyurea	**Positive**	1.6 – 200	7.8 – 500
Ibuprofen	Negative	2.9 – 1400	47 – 3000
Isotretinoin	**Positive**	9.3 – 300	0.0001 – 250
Imatinib	**Positive**	3.1 – 100	0.8 – 50
IWP-2	**Positive**	0.003 – 0.1	0.8 – 50

Lazabemide	Negative	3.1 – 100	0.6 – 400
Metformin	Negative	15.6 – 500	0.7 – 500
Methotrexate	**Positive**	0.003 – 40	0.0001 – 1
Misoprostol	**Positive**	0.04 – 1.3	1.3 – 100
Penicillin G	Negative	4.7 – 600	31 – 2000
Progesterone	Negative	0.63 – 40	0.3 – 500
Retinoic Acid	**Positive**	0.0003 – 0.035	0.00016 – 350
RO-1*	**Positive**	3.1 – 100	3.1 – 300
RO-2*	Negative	15.6 – 500	0.3 – 250
RO-3*	**Positive**	9.4 – 300	3.9 – 500
RO-4*	Negative	6.3 – 200	3.9 – 500
RO-5*	Negative	1.6 – 50	0.07 – 50
RO-6*	Negative	12.5 – 400	0.0003 – 250
RO-7*	Negative	18.8 – 600	0.6 – 400
RO-8*	**Positive**	2.5 – 80	1.3 – 125
RO-9*	**Positive**	0.16 – 5	0.8 – 50
RO-10*	**Positive**	0.5 – 15	0.07 – 50
RO-11*	Negative	1.25 – 40	2.3 – 150
RO-12*	Negative	3.1 – 100	3.9 – 250
SB431542	**Positive**	0.63 – 20	0.1 –100
(±) Thalidomide	**Positive**	0.001 – 0.5	125 – 2000
Valproic Acid	**Positive**	31.25 – 1000	47 –3000
Warfarin	**Positive**	1.9– 60	39– 2500

2.6. 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 differential gene expression data of the first six germ-layer factors into three levels using the Hartemink's pairwise mutual information method implemented in the *bnlearn* package (74). We generated 1,000 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 persist in more than 85% bootstrap samples are deemed as significant and reported.

The beta regression model used for sensitivity analysis was built with the *glmmTMB* package (75). Scores outside the boundaries [0.01, 0.99] are set to the boundary values to allow beta regression. All ten factors and significant interaction terms identified in the Bayesian network are used as the model input, and compounds are modelled as random effects to capture between-concentration correlations. For better interpretability, input variables are scaled to 0 mean and standard deviation. Simulation was performed with the *ggeffects* package (76).

3. Results

3.1. Gene expression quantification by molecular phenotyping

We described previously that differential expression of a set of 87 genes preferentially expressed in different germ layers (*germ-layer genes* hereafter), which both determine and reflect embryonic development, is in principle able to distinguish between teratogenic and non-teratogenic compounds (23,25). To validate our findings, we compiled a large set of well-documented teratogens, partially with label information for drug-use, and non-teratogens that are challenging to predict and/or known to cause false-positives using animal studies (Suppl. Tab. S1, S2). The compounds cover a broad spectrum of chemical classes and a wide range of effective concentrations. This large compilation of compounds with solid clinical and animal data anchoring is a useful resource for further model development.

We interrogated our human stem-cell model with the compilation of compounds, adapting the experimental workflow that we developed previously (Fig. 1a and 1b). We identified the assay throughput as a major challenge due to the high number of samples for gene expression profiling (>1,000). It would be particularly cost- and labor-intensive if we use the digital PCR technique, established in our previous work to quantify gene expression (23). To address this challenge, we used molecular phenotyping as alternative readout. Molecular phenotyping is an amplicon-based targeted sequencing approach, which delivered quantitative expression data of 1,215 detectable genes. Notably, all germ-layer specific genes used in our previous work were included. In this way, we were able to characterize both general pathway activity modulations and germ layer-specific changes as potential features associated with teratogenicity (26-28).

We performed extensive quality control of the data. In particular, we addressed the questions whether results of molecular phenotyping are comparable to those of qRT-PCR, and whether the hiPSC used show expected reproducibility based on their gene expression profile. We compared the differential expression profiles of germ-layer genes obtained by RT-qPCR in previous studies with newly generated data of molecular phenotyping and observed highly similar results (Pearson correlation coefficient R=0.9, p<2.2E16) (Fig. 1c). This suggests that targeted RNA sequencing with molecular phenotyping delivers highly comparable results, at least for germ-layer genes. The comparison is not feasible for other pathway reporter genes because they were not quantified by digital RT-qPCR.

A unique advantage of quantifying pathway reporter genes along with germ-layer genes is that we can use them to assess 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 (31). 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 (Fig. 1d). It suggests that the hiPSC used for the assay shows 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 (23,25).

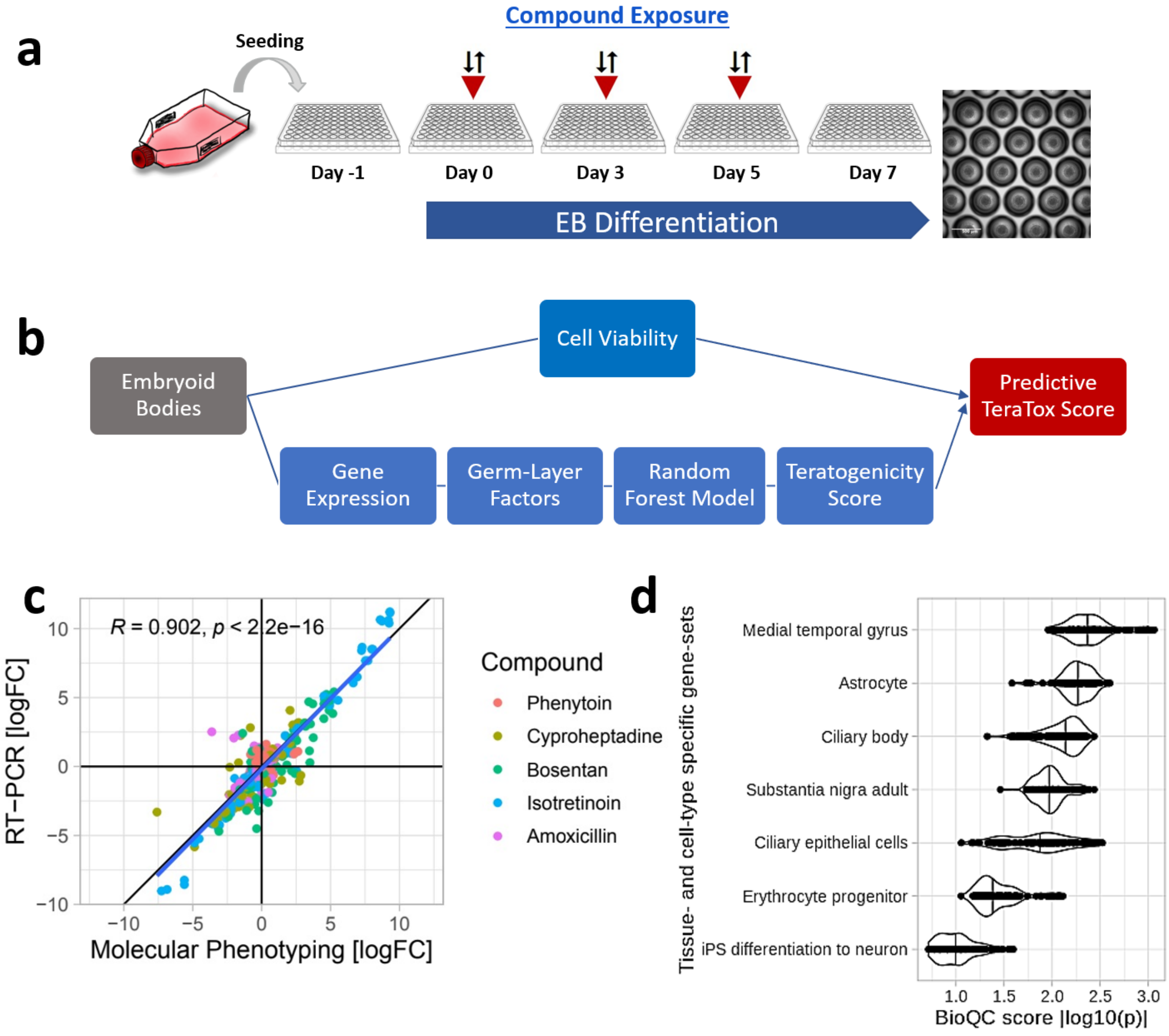


Figure 1: The human *TeraTox* assay: workflow and quality control.

(a) Workflow of the human *TeraTox* assay: cell culture and compound testing. To start, human induced pluripotent stem cells (hiPSC) are seeded in a density of 120.000 cells per well in 96-well microwell plates. They form homogenous embryoid bodies (EBs) which are spontaneously differentiated over 7 days. Compounds are added with six-point concentrations at day 0, day 3 and day 5. Single wells containing about 70 differentiated EBs are lysed to generate one sample for the viability assay and gene expression profiling.

(b) Workflow continued: cell viability and gene expression profiling. After 7 days of spontaneous EB differentiation, cell viability of EBs is determined by ATP release with the luminescent CellTiter-Glo assay, and normalized to DMSO controls by setting the average of latter to 100%. Gene expression was determined with molecular phenotyping by targeted RNA sequencing. We derive differential expression profiles (vectors of z-scores) of germ-layer genes induced by compounds with linear models comparing with DMSO, and summarize them by the profiles of 10 germ-layer factors that we identified with factor analysis (further detailed in Figure 2). Median z-scores of genes belonging to each germ-layer factor are used as the input for a random-forest regression model to predict teratogenicity scores (further detailed in Figure 3). We generate concentration-response curves of both cell viability and teratogenicity scores, and calculate a *TeraTox* Score per compound to predict its teratogenicity potential (further detailed in Figure 4).

(c) Differential gene expression derived from molecular phenotyping in *TeraTox* is highly correlated with differential gene expression derived from the gold-standard RT-PCR assay. The scatter plot shows differential expression (logFC: log2 fold-change) of germ-layer genes derived from targeted RNA sequencing with the molecular phenotyping platform, analyzed by the *edgeR* method, in x-axis, and differential gene expression quantified with RT-qPCR, analyzed by the *ddCt* method (25), in y-axis. Each dot represents a gene. Samples shown are independent biological replicates treated with identical compounds and concentrations, and median values of technical replicates are used for plotting. Colors indicate compounds that were tested with both techniques. Solid lines indicate *x=0*, *y=0*, and *y=x*, respectively. *R* gives the Spearman correlation coefficient between the two sets of measurements across compounds, and the *P*-value reports the probability that the null hypothesis (no correlation between the two measurements) is rejected wrongly.

(d) Gene expression profiles of hiPSC used by *TeraTox* reveal their biological identity. We applied *BioQC* analysis to raw gene expression data from all 1,055 samples. BioQC identifies enrichment of characteristic expression signatures from tissues, organs, and cell types to assess sample-specific constitution. The gene-sets are derived from large gene expression compendia. BioQC scores are absolute log10 transformed *p-values* of the Wilcoxon-Mann-Whitney test. The larger the score, the more enriched is the expression of the set of genes of interest *i.e.,* the expression of genes in the set are higher than genes that are not in the set. Each dot represents one sample. Violins indicate the distributions of the BioQC scores of each gene set, respectively, with vertical lines indicate median values.

3.2. Unsupervised learning from gene expression data with factor analysis

Before applying supervised learning techniques to differentiate teratogens from non-teratogens, we applied several unsupervised learning algorithms to analyze 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 differential gene expression (data not shown). Unexpectedly, factor analysis revealed both biological insights and, as further discussed below, a feature engineering technique that contributed to the best-performing model. Since this is the first time to our knowledge that factor analysis is applied in the context of gene-expression-based toxicity prediction, we highlight its concepts and unique advantages.

Factor analysis, sometimes called exploratory factor analysis to differentiate it from confirmatory factor analysis, is a statistical method to discover latent (unobserved) variables that account for the correlations observed between features. Useful for both dimension reduction and feature engineering, factor analysis has been particularly powerful in building predictive machine-learning models in biology using highly correlated features such as cell morphology in the context of high-content screening (32,33,77,78). With respect to gene expression, factor analysis reduces the data dimension from genes to factors, each of which is usually associated with multiple genes. Genes in each factor show correlated gene expression profiles across samples (Fig. 2a, b). These factors, therefore, can be thought of as being a representation of all biological processes influencing gene expression, for instance epigenetic profiles, transcription factor activities, microRNA abundances, etc. Despite the fact that most of these variables are not directly observable, latent factor analysis offers a possibility to infer their total contribution to detected variation in gene expression profiles.

Conceptually, factor analysis is familiar with other correlation-based methods, for instance Relevance Networks (79) and Weighted Correlation Network Analysis (WGCNA) (80). We preferred factor analysis to alternative methods because factor analysis does not make any additional assumptions than the common, minimum ones underlying correlation analyzes (homogeneity, completeness, *etc.*), whereas other methods do so, for instance the scale-free network structure assumed by WGCNA, whereas this assumption is often challenged (81,82). On the other hand, we have many more samples than the number of factors. Factor analysis is feasible with the maximum-likelihood method. We therefore decided to use factor analysis following the principle of Occam's Razor.

We applied factor analysis to raw gene expression data and identified intriguing patterns. Since factor analysis is based on inter-gene correlations, we visualize the correlation matrix of germ-layer genes in Figure 2a (the full matrix is visualized in Supplementary Figure S2a). 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, it revealed biologically meaningful patterns. Using the maximum likelihood method, we decomposed the covariance matrix of gene expression into factors. The heatmap in Figure 2b shows loadings, *i.e.* how strong factors influence the expression of germ-layer genes, of the first ten factors that collectively explain more than 70% of the covariance (Suppl. Fig. S2b and S2c). Left to the heatmap we use colors to indicate germ-layer classes that were distilled from biological knowledge. We found that the first six factors (ranked by explained covariance of the data) are significantly enriched with signatures of individual germ layers or signatures of stem-cell self-renewal (Fig. 2c, p<0.01, Fisher's exact test). This significant enrichment is both intriguing and novel, because while it is established that germ-layer genes are highly expressed at different stages of embryogenesis, we failed to find any previous studies reporting that their expression are strongly correlated in 3D embryoid bodies formed by hiPSC, with or without compound treatment. Given that the cells in *TeraTox* are all grown 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 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, *e.g.*, *WNT1*, *POU4F1*, *OLFM3*, *CDH9*, *LMX1A*, *DMBX1*, *PAX3*, *MAP2*, and *TRPM8* (Fig. 2a). While BioQC analysis revealed that ectodermal genes are highly expressed at the endpoint on day 7, factor analysis further indicates 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 (Factor 4-6), respectively. The remaining factors (Factor 7-10) are of smaller sizes and more heterogeneous (Fig. 2b). Genes associated with each factor are associated mainly, but not exclusively, with other genes of the same germ-layer class.

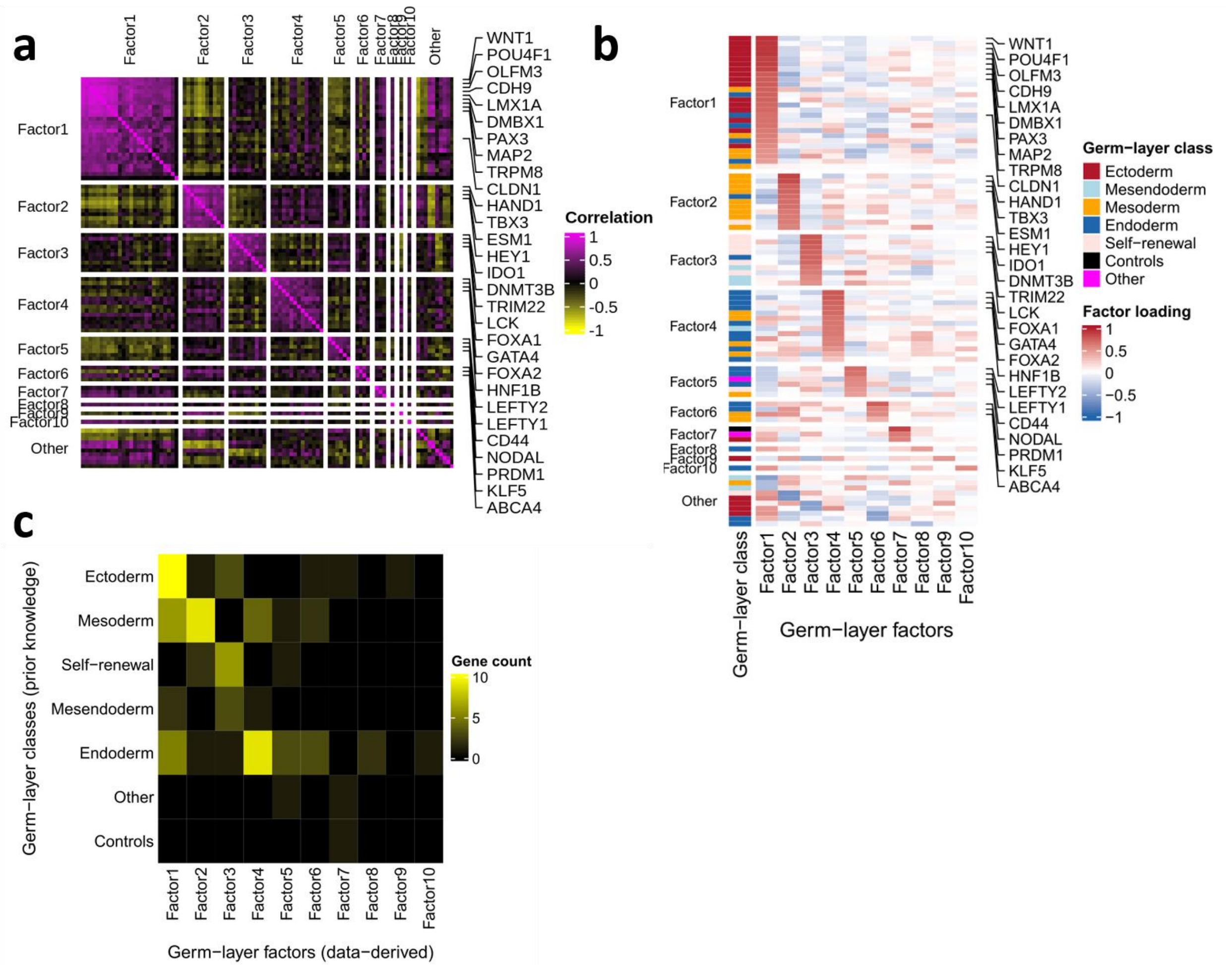


Figure 2: Identification of latent factors that are associated with germ layers.

(a) Germ-layer genes correlate with each other and form clusters. The heatmap represents Pearson correlation coefficients of gene expression in all samples, including both vehicle controls and treated samples. Each row and each column represent a gene and the matrix is symmetric. Labels are shown for representative genes from each cluster of strongly correlated genes (the full matrix is shown in Supplementary Figure S2a). To assist interpretation of the latent germ-layer factors, genes are split by them. Magenta colors represent strong positive correlations between genes, yellow colors represent strong negative correlations, and black colors represent little correlation.

(b) Loadings of germ-layer factors. The heatmap shows loadings of latent factors on the germ-layer genes. A loading equal to or near 1 indicates that the factor strongly influences the gene. A loading near 0 means that the factor has little effect on the gene. And a loading equal to or near -1 indicates that the factor negatively influences the gene. Each row contains a germ-lay gene, and each column contains a factor. The rows are ordered so that the genes that are ordered by the factors are impacted most, which is visible from the patterns of cascades. Though no prior biological knowledge is used in the analysis, we noticed that the factor loadings partially resemble the clustering of genes defined by germ-layer classes, which are illustrated in the row-side color to the left of the figure. Based on this reason, we call the factors *germ-layer factors*. We consider 77 out of 87 germ-layer genes for germ-layer factors because the remaining 10 genes are negatively influenced by germ-layer factors (bottom rows in the heatmap), and therefore it does not make sense averaging their expression with other genes that are positively influenced. For readability, we showed representative genes from the first six clusters here while displaying the full matrix in Suppl. Fig. S2c.

(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).

3.3. 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 as input, we explored all combinations of the following options exhaustively (Fig. 3a):

1. *Feature type*: We tested both log2 fold change (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 differential gene expression.
2. *Feature engineering*: We used all detectable pathway reporter genes (N=1,215), 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). In case of both germ-layer classes and factors, we use the median value of the genes belonging to each group as the engineered feature.
3. *Model construction*: We used and benchmarked two methods of different nature, Elastic Net (linear regression with regularization) and Random Forest (ensemble decision trees), to construct machine-learning models. We chose them based on the size of the dataset and the relatively good explainability of both methods (83).

4. *Target variable*: We used both binary classification (teratogen or non-teratogen) and regression (the teratogenicity score, defined below and further detailed in the Material 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 leave-one-out (LOO) scheme. In the first case, we used 80% compounds (stratified sampling from non-teratogens and teratogens) as the training set to train a model, which was used to predict the teratogenicity scores using the remaining 20% compounds as the test set. In the latter case, all except one compound were used to train the model, which predicts the teratogenicity scores for the left-out compound, and repeated the procedure for all compounds so that teratogenicity scores were predicted for each compound based on data from other compounds. In either case, the model performance was assessed by F_1 scores in case of binary classification models, and Spearman correlation coefficients of teratogenicity scores for teratogens in case of regression models. The best model parameters were searched by 10-fold cross-validations of the training set.

While all other technical terms are used in their common sense, we explain the motivation and definition of the Teratogenicity Score in detail. A key challenge for building a predictive model of teratogenicity is that the potential of a compound inducing teratogenicity varies by its concentration. A concentration-response relationship can be assumed, namely a treatment with a higher concentration is more likely to induce teratogenicity than that with a low concentration. However, the concrete functional form between the potential and the concentration is not known. This motivated us to define the Teratogenicity Score as the '0-1 cosine bounded similarity' between differential gene expression profiles induced by any given concentration and the profiles induced by the maximum non-cytotoxic concentration. The teratogenicity scores of teratogens are defined between 0 and 1, and those of non-teratogens are fixed as 0 at all concentrations (Fig. 3b). By defining teratogenicity scores, we effectively transformed the binary classification problem into a regression problem.

Two important technical details require clarification. First is the range of the teratogenicity score. Mathematically, cosine similarity ranges between -1 and 1; we bounded it to 0-1 by setting negative similarities as zero, which did not change the performance of the models (data not shown) but helped with human understanding. The teratogenicity score can be interpreted as an estimate of the probability of inducing teratogenicity, which would be a real number between 0 and 1, though the real probability is unknown to us because we are working with an *in vitro* system only, and the probability estimated in our system may differ significantly from that *in vivo*.

The second technical detail is the selection of regression models. Given the truncated domain where the teratogenicity score is defined, we tried both simple linear regression and generalized linear models with beta regression. However, beta regression was computationally intensive and much slower, and its use led to similar results as simple linear regression for predicting teratogenicity scores. Therefore, we used simple linear regression throughout the study except in the last part of model explainability, because only one model is required there and the boundary consideration is important for simulation studies.

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 (Fig. 3c and 3d, contrasted with Fig. 5a). For teratogenicity prediction, the combination of germ-layer factors and random-forest regression works the best.
3. With regard to the *target variable*, the performance of the regression-based teratogenicity-score prediction model is slightly better than the model for binary classification (data not shown).

4. Performance is comparable between two modes of *data splitting* (data not shown)*.* However, the leave-one-out training-testing scheme is preferable because it allows us to set up a single threshold of teratogenicity score which can be applied to all compounds, 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 teratogenicity score as the target variable to build the predictive model for teratogenicity with gene expression data.

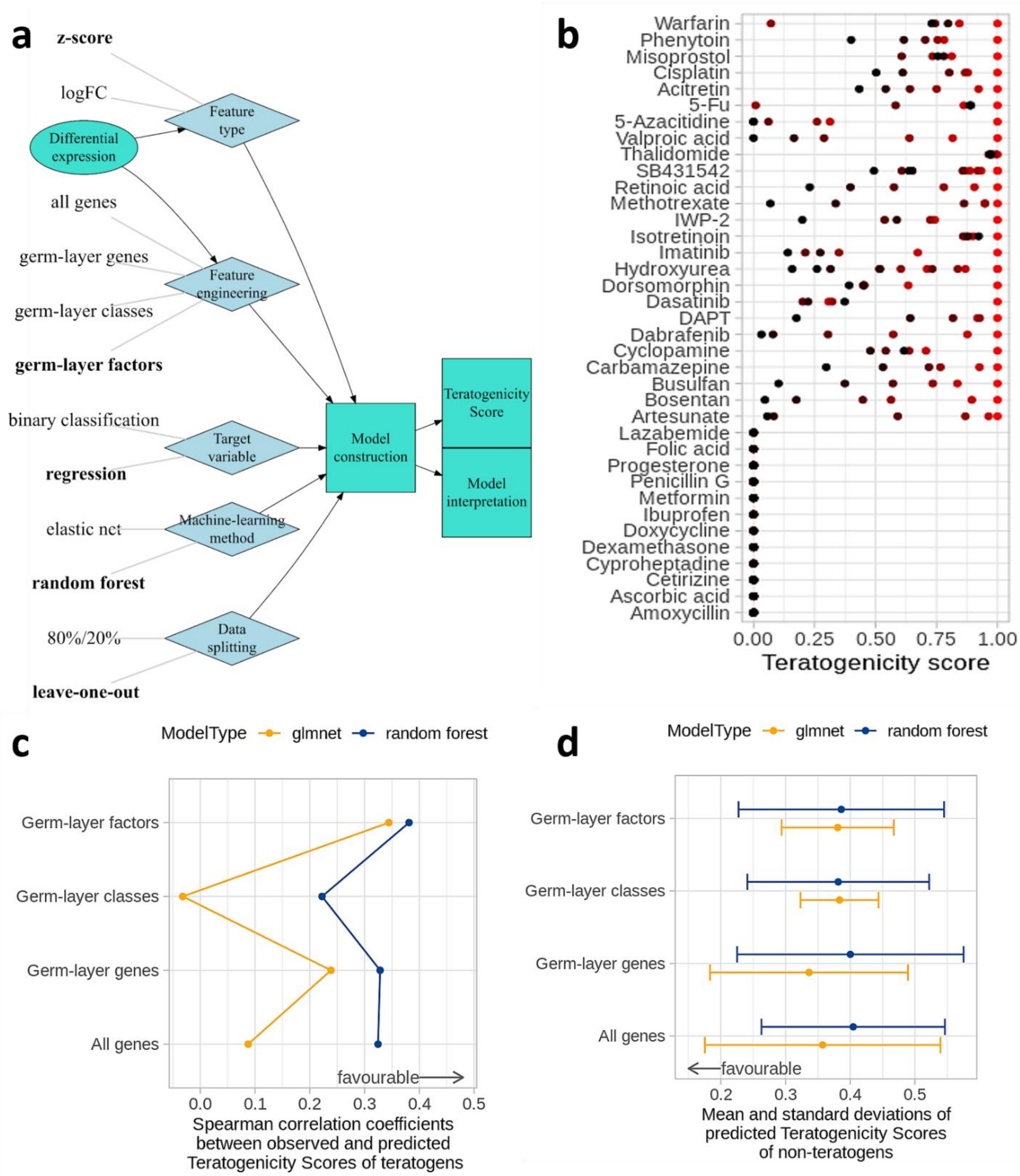

a
z-score
logFC
Differential expression
Feature type
all genes
germ-layer genes
germ-layer classes
germ-layer factors
Feature engineering
binary classification
regression
Target variable
elastic net
random forest
Machine-learning method
80%/20%
leave-one-out
Data splitting
Model construction
Teratogenicity Score
Model interpretation
b
Warfarin
Phenytoin
Misoprostol
Cisplatin
Acitretin
5-Fu
5-Azacitidine
Valproic acid
Thalidomide
SB431542
Retinoic acid
Methotrexate
IWP-2
Isotretinoin
Imatinib
Hydroxyurea
Dorsomorphin
Dasatinib
DAPT
Dabrafenib
Cyclopamine
Carbamazepine
Busulfan
Bosentan
Artesunate
Lazabemide
Folic acid
Progesterone
Penicillin G
Metformin
Ibuprofen
Doxycycline
Dexamethasone
Cyproheptadine
Cetirizine
Ascorbic acid
Amoxycillin
0.00
0.25
0.50
0.75
1.00
Teratogenicity score
c
ModelType
glmnet
random forest
Germ-layer factors
Germ-layer classes
Germ-layer genes
All genes
favourable
0.0
0.1
0.2
0.3
0.4
0.5
Spearman correlation coefficients between observed and predicted Teratogenicity Scores of teratogens
d
ModelType
glmnet
random forest
Germ-layer factors
Germ-layer classes
Germ-layer genes
All genes
favourable
0.2
0.3
0.4
0.5
Mean and standard deviations of predicted Teratogenicity Scores of non-teratogens

Figure 3: Construction of machine-learning models predicting concentration-dependent teratogenicity potentials based on differential gene expression as input.

(a) Overview of the workflow to construct machine-learning models using differential gene expression as input to predict teratogenicity potential. Oval node: input; diamond nodes: steps where more than one option was tested; plain-text nodes: options that were tested; rectangle nodes: outcome of the model. Plain-text nodes in bold show the options that give the best prediction performance. logFC: log2 fold change.

(b) Definition of teratogenicity scores (TS). TS is set to 0 for non-teratogens (from Amoxicillin to Lazabemide from bottom up), independent of the concentration level. For teratogens, TS is set to 1 for the highest non-cytotoxic concentration, and TS for other concentrations is set to the cosine similarity of differential gene expression profiles between each concentration and the highest non-cytotoxic concentration. Negative TS is set to 0. Colors indicate the concentration level: the highest concentration is assigned red, the lowest concentration is assigned black, and the concentrations between them are of darker red as they move towards lower concentrations. Points for non-teratogens are overlapping with each other. Here we show the subset of commercially available compounds as examples. The same definition was also applied to proprietary compounds.

(c) Spearman correlation coefficients between observed teratogenicity scores, calculated on a per-compound basis, and predicted teratogenicity scores, which are derived from models trained using all but the test compound (leave-one-out). Higher values are favorable. Only teratogens are considered here because the teratogenicity scores are defined as 0 for non-teratogens and the Spearman correlation coefficient is a poor choice to characterize the models in such cases.

(d) Mean (dots) and standard deviations (error bars) of teratogenicity scores of non-teratogens. For each compound, the median teratogenicity score is derived from six concentrations to represent the compound. Both lower values and smaller error bars are favorable.

3.4. Assay performance of the *TeraTox* assay compared to mEST prediction

Based on the best-performing machine-learning model, we defined the following predictive model for teratogenicity. First, we considered the maximal non-cytotoxic threshold concentration (NCC_{max}) for cell viability measured by the CellTiter Glo assay of at least 80%. Next, we defined the minimal teratogenic concentration (TC_{min}) as the concentration at which the threshold of the teratogenicity score was met (TS=0.38, defined by grid search, Fig. 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* to avoid confusion with *Teratogenicity 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 (Fig. 4b).

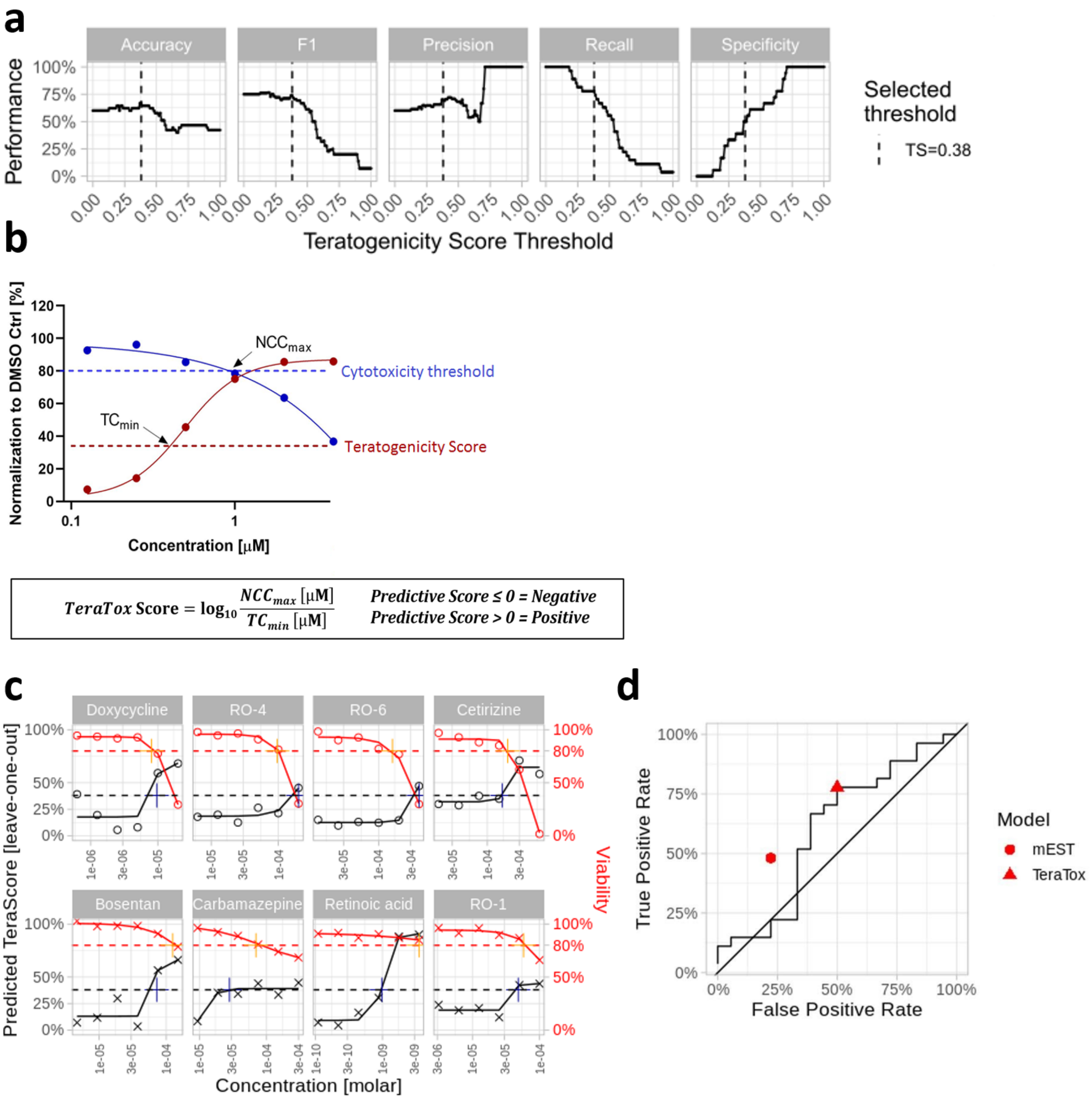
We plotted the concentration-response curves of both measured cytotoxicity and predicted teratogenicity scores induced by each compound (Fig. 4c, see Suppl. Fig. S4 for all compounds). In general, teratogenicity levels increased while cell viability decreased with rising concentrations. Correctly predicted negative compounds were unlikely to induce teratogenicity within non-cytotoxic concentrations, which means the calculated *TeraTox* score was negative or zero (e.g., Doxycycline, RO-4, RO-6). Positive compounds (e.g., Bosentan, Carbamazepine, Retinoic Acid, RO-1) or false positive predicted compounds (e.g., Cetirizine) were more likely to induce teratogenicity under non-cytotoxic concentrations, which was indicated by positive *TeraTox* scores (Fig. 4c).

We compared predictions of 45 reference compounds by *TeraTox* scores with classifications from FDA or *in vivo* EFD studies (Suppl. Tab. S4). Classification with *TeraTox* Scores achieved an overall accuracy of 68% and outperformed mEST (60%). The two assays show different sensitivity and specificity profiles: While mEST is more specific (specificity 78%), *TeraTox* is more sensitive (sensitivity/recall 78%). Among 18 negative reference compounds, 9 were classified as false positives (FP) by TeraTox, and only 4 by the mEST. Whereas from 27 positive reference compounds, 21 were predicted as true positives (TP) by the human *TeraTox* and only 13 by the mEST (Tab. 2, Fig. 4d). It is noteworthy that among the 26 compounds misclassified in total, these seven are wrongly predicted by both assays: cyproheptadine, RO-11, 5-FU, methotrexate, misoprostol, RO-8, warfarin. Given the distinct sensitivity and specificity profiles of the two assays, we asked whether we can achieve even better prediction results by using them in a sequential mode. Specifically, we first let mEST classify the compounds, and among the negative predictions, we accept the predictions by *TeraTox*. The intuition is that we may benefit both 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%. This suggests that it may be possible to achieve better prediction results by combining the existing mEST assay with the novel *TeraTox* assay.

Table 2: Overview of assay performance for mEST and human *TeraTox* assay.

Values were calculated based on 45 compounds (according to equations 1-5 in section 2.4. TP= true positive, TN=true negative, FP=false positive, FN=false negative).

Model	***TP***	***TN***	***FP***	***FN***	***Accuracy***	***Precision***	***Recall***	***Specificity***	***F_1***
TeraTox	21	9	9	6	67%	70%	78%	50%	73%
mEST	13	14	4	14	60%	76%	48%	78%	59%
mEST + TeraTox	21	14	4	6	78%	84%	78%	78%	81%



$$\textit{TeraTox}\ \text{Score} = \log_{10}\frac{NCC_{max}\,[\mu\text{M}]}{TC_{min}\,[\mu\text{M}]}$$

Predictive Score ≤ 0 = Negative
Predictive Score > 0 = Positive

Figure 4: Prediction of teratogenicity with the human *TeraTox* assay.

(a) Results of a grid search to select the optimal threshold of the teratogenicity score. Prior to the grid search, we predicted teratogenicity scores for each compound using data from all other compounds (with the best options, *i.e*. using z-scores, germ-layer factors, random-forest, regression-based prediction, and leave-one-out cross-validation). We then fitted concentration-response curves to the predicted teratogenicity scores and made predictions by the model described in Figure 4b by varying the thresholds in a grid search. Each dot in the plot indicates one point in the grid, which starts at 0 and ends at 1, with a step size of 0.01. The best threshold (TS=0.38) was chosen manually by inspecting the performance metrics defined in Equations (1)-(5).

(b) Concentration-response curves normalized to DMSO solvent controls for determination of minimal teratogenic concentration (TC_{min}) and maximal non-cytotoxic concentration (NCC_{max}) using predicted teratogenicity scores and measured cell viability. Teratogenicity scores were predicted by leave-one-out testing/ training, and the optimal threshold of teratogenicity scores (TS=0.38) was found by grid search (a). The log ratio of the concentration leading to 20% viability impairment (NCC_{max}) and that causing a teratogenicity score equal to the threshold (TC_{min}) was used to calculate a predictive score. If no NCC_{max} or TC_{min} could be determined, the minimum tested concentrations were used for TC_{min} and the maximum tested concentrations were chosen for NCC_{max}. Predictive scores ≤0 classified the compounds as negative and values >0 were classified as positive.

(c) Examples of concentration-response curves reported by the *TeraTox* assay of 4 selected non-teratogens (top panel, concentrations indicated by open circles): Doxycycline, RO-4, RO-6, and Cetirizine, and 4 selected teratogens (bottom panel, concentrations indicated by crosses): Bosentan, Carbamazepine, Retinoic Acid, and RO-1. In most cases, teratogenicity (black curves) rises with increasing concentrations whereas cell viability (red curves) decreases. The points indicate predicted teratogenicity score (2 replicates each) and measured cytotoxicity (3 replicates each).

(d) The receiver operating characteristics (ROC) curve based on 45 reference compounds. The triangle symbol indicates the performance of *TeraTox*, and the circle indicates the performance of mEST.

3.5. Model interpretation and explanation

A model's explainability is crucial for understanding that allows inspection and further improvement (84). We performed additional in-depth analysis and collected data orthogonal to *TeraTox*, thereby implementing three independent approaches to interpret and explain how the *TeraTox* model, in particular, how the teratogenicity score prediction model works.

First, we followed up on previous work and asked the question whether the cytotoxicity quantified by the phenotypic assay can be predicted by gene expression data as well, and whether teratogenicity scores are confounded by general cytotoxicity (85,86). For this purpose, we followed the same scheme as described in Figure 3a while using cytotoxicity instead of teratogenicity scores as the target variable. Interestingly, an exhaustive 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, gives the best result (Fig. 5a).

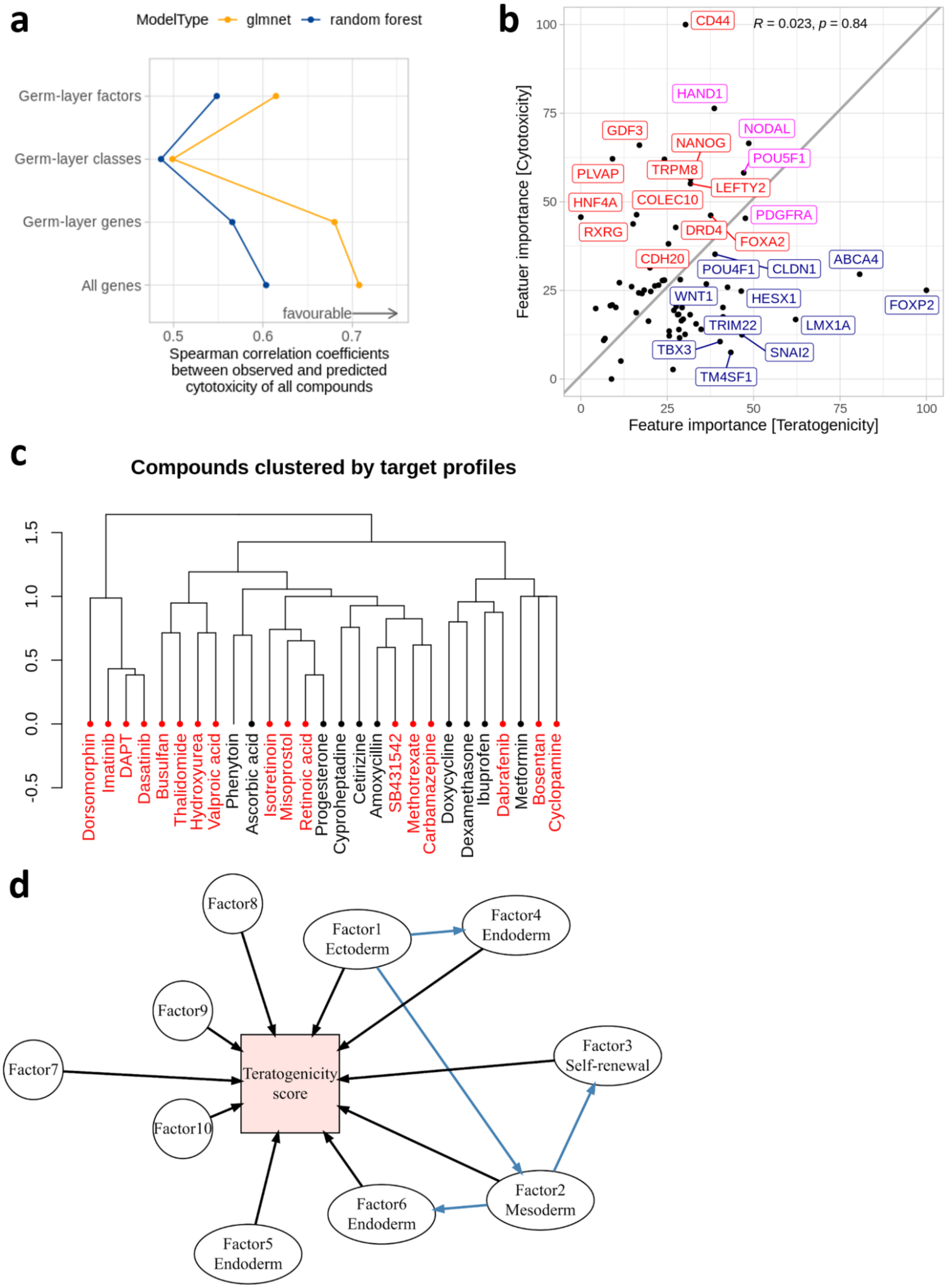
Given that the combination of germ-layer genes and random forest gives reasonable performance in both cases, and that random forest allows inquiry of feature importance by accuracy, we compared the feature importance of germ-layer genes in predicting both target variables (Fig. 5b). The prediction of cytotoxicity and teratogenicity by molecular phenotyping relies on expression changes of distinct genes. The distinction shows that teratogenicity of a compound is not a determinant for cytotoxicity whereas a compound that shows cytotoxicity at a specific concentration can still be teratogenic at lower, non-cytotoxic concentrations and that pathways for cytotoxicity and teratogenicity may be independently regulated. This is well in line with several previous findings (87-89).

The second approach addressed the question whether a compound's pharmacology, namely 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. While some teratogens indeed have similar target profiles, we observe close clustering of teratogens and non-teratogens that have similar target profiles as well (Fig. 5c, Suppl. Fig. S5a). 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 differential gene expression across concentration per compound and that of pharmacological profiles (Suppl. Fig. S5b). 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 predict a compound's teratogenicity potential, at least in their current stand. *In-vitro* assays, for instance with *TeraTox* and other advanced cellular models*,* are indispensable for preclinical teratogenicity assessment.

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 beta 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 (DAG). Under this assumption, we built a Bayesian network using the differential expression data of germ-layer factors (Fig. 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 (Fig. 5e, Suppl. Fig. S6). The model provides both interpretable coefficients of the model and a tool for sensitivity analysis, because we can quantify prediction uncertainty much easier with a linear model than the 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 teratogenicity scores. 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 (Fig. 5e). The results of sensitivity analysis further underlined the prominent ectodermal nature of the model at the endpoint on day 7.

In summary, we explain how the *TeraTox* model works by complementing the machine-learning model with feature importance analysis, biological and pharmacological interpretation, and sensitivity analysis.


a
ModelType
glmnet
random forest
Germ-layer factors
Germ-layer classes
Germ-layer genes
All genes
favourable
0.5
0.6
0.7
Spearman correlation coefficients between observed and predicted cytotoxicity of all compounds
b
R = 0.023, p = 0.84
Featuer importance [Cytotoxicity]
Feature importance [Teratogenicity]
0
25
50
75
100
CD44
HAND1
GDF3
NODAL
NANOG
POU5F1
PLVAP
TRPM8
LEFTY2
HNF4A
COLEC10
PDGFRA
RXRG
DRD4
FOXA2
CDH20
POU4F1
CLDN1
ABCA4
WNT1
HESX1
FOXP2
TRIM22
LMX1A
TBX3
SNAI2
TM4SF1
c
Compounds clustered by target profiles
-0.5
0.0
0.5
1.0
1.5
Dorsomorphin
Imatinib
DAPT
Dasatinib
Busulfan
Thalidomide
Hydroxyurea
Valproic acid
Phenytoin
Ascorbic acid
Isotretinoin
Misoprostol
Retinoic acid
Progesterone
Cyproheptadine
Cetirizine
Amoxycillin
SB431542
Methotrexate
Carbamazepine
Doxycycline
Dexamethasone
Ibuprofen
Dabrafenib
Metformin
Bosentan
Cyclopamine
d
Factor8
Factor1 Ectoderm
Factor4 Endoderm
Factor9
Factor7
Teratogenicity score
Factor3 Self-renewal
Factor10
Factor6 Endoderm
Factor2 Mesoderm
Factor5 Endoderm

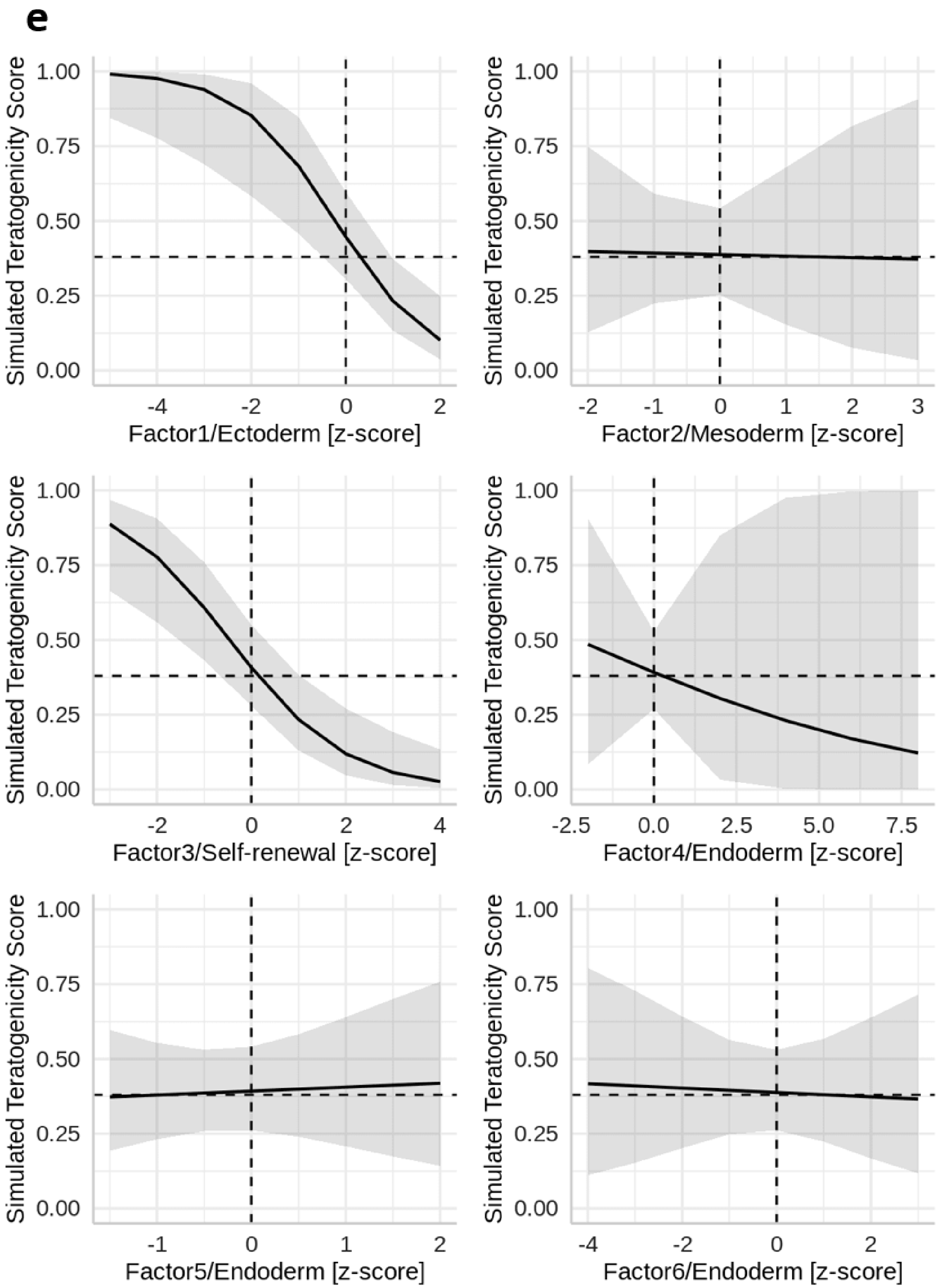

e
Simulated Teratogenicity Score
1.00
0.75
0.50
0.25
0.00
-4
-2
0
2
Factor1/Ectoderm [z-score]
Simulated Teratogenicity Score
-2
-1
0
1
2
3
Factor2/Mesoderm [z-score]
Simulated Teratogenicity Score
-2
0
2
4
Factor3/Self-renewal [z-score]
Simulated Teratogenicity Score
-2.5
0.0
2.5
5.0
7.5
Factor4/Endoderm [z-score]
Simulated Teratogenicity Score
-1
0
1
2
Factor5/Endoderm [z-score]
Simulated Teratogenicity Score
-4
-2
0
2
Factor6/Endoderm [z-score]

Figure 5: Biological interpretation of the model.

(a) Differential gene expression can be used to predict cytotoxicity, though the best-performing model for cytotoxicity differs from that for teratogenicity. Similar as what we did for teratogenicity in Figure 3b, we built machine-learning models for cytotoxicity using the same workflow shown in Figure 3a. While z-scores, regression, and leave-one-out remain the same, the best model for cytotoxicity predictions uses all pathway reporter genes and elastic net, in contrast to germ-layer factors and random forest for teratogenicity prediction. Each dot represents the correlation between held-out and predicted cytotoxicity and observed cytotoxicity using leave-one-out cross validation. Lines of two colors (blue: random forest; orange: elastic net, also known as *glmnet*) show that the elastic net consistently outperforms random forest for cytotoxicity prediction.

(b) Feature importance of germ-layer genes differ for cytotoxicity and teratogenicity prediction. Because the combination of germ-layer genes and random forest shows reasonable predictivity for both teratogenicity (Fig. 3b) and cytotoxicity (Fig. 5a), we inspected their feature importance in the random forest model using importance measures. Each dot represents a germ-layer gene. Genes that are mostly important for teratogenicity prediction are shown in blue. Genes that are most important for cytotoxicity prediction are shown in red. The overlapping genes that are important for both predictions are shown in magenta.

(c) The target profile alone is not sufficient to determine a compound's teratogenicity potential. We clustered compounds by their target profile, *i.e.* number and affinity to targets collected in the ChEMBL database and visualized the clustering with the dendrogram. Compounds are colored by their truth classification whether they are teratogen (red) or not (black). Note that while teratogens are enriched in some branches of the dendrogram, non-teratogens and teratogens can have very similar target profiles and therefore cluster near to each other in other cases.

(d) Structure of the Directed Acyclic Graph (DAG) that we used to model the relationship between teratogenicity score and germ-layer factors with generalized linear model. Besides ten germ-layer factors, significant interactions between germ-layers identified by Bayesian networks (blue edges) are used as input variables for the model. Model fitting results are shown in Supplementary Figure S6.

(e) Sensitivity analysis shows both the advantages and the limitations of the *TeraTox* assay. The relatively simple generalized linear model with beta-regression allowed us to run the sensitivity analysis, a simulation technique to test how the model would behave if we tune the input variables specifically. Each panel shows one of such analyzes, where we tune one parameter (for instance the ectoderm germ-layer factor in the top-left panel) while keeping all other parameters fixed. Black lines indicate average prediction and gray areas indicate 95% confidence intervals of prediction. To facilitate interpretation, the input variables are scaled to 0 mean and standard deviation. Note that for all sensitivity analyzes plots, when the input parameter is 0 (vertical dashes), the simulated teratogenicity scores center around the optimal threshold that we identified (TS=0.38, horizontal dashes). Therefore, the plot can be interpreted in the following way: if the slope of a variable is positive, the teratogenicity potential increases as the expression of genes in that germ-layer factor increases. Otherwise, if the slope is negative, the teratogenicity potential increases as the expression of genes in that germ-layer factor decreases.

4. Discussions

This study characterizes the optimization of *TeraTox*, a newly developed human teratogenicity assay. *TeraTox* quantifies drug-like molecules' cytotoxicity and teratogenicity profiles in concentration response using a hiPSC derived embryoid body model that spontaneously differentiates into all three germ layers over seven days. It thus extended and standardized earlier embryoid body models, and fully leveraged their predictive potential by adding a toxicological prediction model (87,90). We challenged the *TeraTox* assay with a selection of 45 reference substances with teratogenic profiles based on high-quality data. We identified latent germ-layer factors that influence germ-layer gene expression, and identified the best machine-learning model that predicts the teratogenicity potential based on germ-layer factors as input and random forest as the regression model. We demonstrated that *TeraTox* outperforms mEST in both sensitivity and balanced prediction performance, though having lower specificity. Furthermore, we explored the interpretation and explainability of the *TeraTox* model with three independent approaches. We found that teratogenicity can be distinguished from cytotoxicity, that pharmacological profiles are not sufficient for predicting teratogenicity, and that the *TeraTox* assay is particularly sensitive towards teratogens impairing ectoderm development and stem-cell self-renewal. The study embodies a comprehensive and critical assessment of the *TeraTox* assay and its predictive algorithm, addressing important open questions for its practical use.

The *TeraTox* model presents a promising companion and an alternative to mEST as a humanized *in vitro* model for preclinical teratogenicity assessment. The two assays differ in cellular origin (human iPSC versus mouse ESC and fibroblasts), final endpoints (differential gene expression from all germ layers versus direct differentiation into mouse cardiomyocytes), and the prediction model. Both assays are anchored to a specific cytotoxicity threshold that determines the non-cytotoxic yet teratogenic effects. Contrary to the mEST assay, where cytotoxicity is inferred from IC_{50} valuesof D3 and 3T3 cells that are grown in monolayers, we anchored the *TeraTox* assay to a much lower cytotoxicity threshold (NCC_{max}, viability >80%) in a three-dimensional scale, which is more physiological relevant. With the exception of a few compounds, *TeraTox* determined cytotoxicity and/or teratogenicity LOAEL (lowest observed adverse effect levels) at lower concentrations compared to the mEST (except of dexamethasone, bosentan, dorsomorphin, hydroxyurea, imatinib, isotretinoin). We therefore believe that *TeraTox* may be a more relevant *in vitro* assay for human teratogenicity assessment.

Our analysis of the *TeraTox* data revealed its three unique advantages over mEST. First, *TeraTox* is more sensitive than the mEST assay. We believe the higher sensitivity is due to several factors, including the use of human induced pluripotent stem cells, cytotoxicity determination in 3D EBs and using gene expression as readout. In this study, we carefully selected concentration ranges based on drug-specific maximum plasma concentrations (C_{max}) from either human data whenever possible or model species otherwise (Suppl. Tab. S1, S2). Retrospective comparison of the *TeraTox* readout with the human therapeutic C_{max} data showed that *TeraTox* captured relevant *in vivo* doses for teratogenicity for most compounds, except for bosentan, isotretinoin, imatinib, and warfarin. The higher sensitivity to detect teratogens is particularly important for preclinical drug discovery to remove potential teratogens from the pipeline as early as possible.

The second advantage of *TeraTox* over mEST is that it allows the detection of human-specific teratogens. Generally, using a model species such as the mouse or the rabbit to predict toxicity may lead to misclassifications if the toxicity is specific for either the species or for humans. For this reason, when we compiled our compound panel, we chose preferentially those compounds that are either known to be species-specific or known to be misclassified by alternative methods. And when we assigned labels to the compounds, we relied on human data whenever possible. An example for species-specific teratogenicity is thalidomide, which was correctly identified as positive by *TeraTox*. At the same time, it shows a high level of cytotoxicity at concentrations that are 80-fold lower than human C_{max}. It is well established that the mouse system is insensitive to the teratogenic effects of thalidomide due to the lack of cereblon-mediated degradation of the *SALL4* transcription factor, which has been shown to result in agenesis of the limb buds in rabbit embryos and was recapitulated by a species-specific false-negative response of the mEST (55,91,92).

The third advantage of *TeraTox* is that it is less of a phenotypic black box but more an interpretable and explainable model. We used factor analysis, an established unsupervised, generative data-analysis method, to reveal clustering patterns in correlations between expression of germ-layer genes. Despite that these clustering patterns, which we termed germ-layer factors, were derived from the raw gene expression data statistically without any biological prior knowledge, we were surprised that they correlated well with known biology of germ-layer development. Specifically, germ-layer factors were enriched with genes preferentially expressed in one of the three germ layers or stem-cell renewal. Interestingly, averaging differential gene expression of germ-layer genes by germ-layer factors provided the best features for the prediction of teratogenicity. The latent factors can be seen as a sum of the output of gene regulatory networks in germ-layer development and stem-cell self-renewal. Therefore, *TeraTox* informs predictions not only based on statistical data patterns: it builds upon biological mechanisms and thus may reflect disturbed functionalities, similar to those leading to teratogenicity *in vivo*. This feature puts the *TeraTox* conceptually in a group of other assays that use phenotypic changes or disturbed functionalities as readouts (17,93-95). The model consolidates our previous call to 'focus on germ layers' and corroborates our recent work exploring gastruloid models that profiles morphological changes of germ-layers for teratogenicity prediction (23,96).

Besides factor analysis, we tried several ways to shed light on how the model works (or not). Most importantly, we could distinguish cytotoxicity from teratogenicity. We explored machine-learning model variants for both teratogenicity and cytotoxicity predictions and made the intriguing observation that the best models are distinctly depending 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. We speculate that there might be two explanations for this. 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 (26). Therefore, we can anticipate that these genes are used by linear regression to predict cytotoxicity. Second, teratogenicity is notably complex. It can be caused in many different subtle ways, with many 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 as summarized by germ layers is probably a more robust readout than individual genes, and random-forest, which is an ensemble learning method, is better at detecting heterogenous signals than linear regression.

Furthermore, we used pharmacological data to show that knowing target profiles of drug candidates is likely not sufficient to predict its teratogenicity potential, therefore an *in vitro* based assay like *TeraTox* is necessary. Last but least, we combined Bayesian network analysis, beta linear regression, and sensitivity analysis to show that while *TeraTox* is sensitive to ectoderm development damage, further work is required to better model mesoderm and endoderm development.

Given the advantages of *TeraTox* over mEST, and considering distinct profiles of sensitivity and specificity of the two assays, we can image three possible scenarios of their routine use in drug discovery: *TeraTox* replacing mEST, *TeraTox* running besides mEST, or two assays running sequentially. We believe while the first option is the long-term goal that we go after, the last option of running them sequentially may be currently the best solution. Our analysis showed that if we use the mEST assay first, and next run the *TeraTox* assay for compounds predicted negative by mEST, we gain improved prediction accuracy, sensitivity, and specificity. Further real-world testing is planned to validate the performance of this approach.

Further studies are warranted to explore several parallel paths further optimizing the *TeraTox* assay, which can be divided into three categories: paths leading to better characterization of EBs, paths leading to better predictive and explanatory algorithms, and paths leading to better biological models of human embryo development. To better characterize EBs, one apparent way is to perform multi-modal - bulk and single-cell omics, and morphological profiling - characterizations of the EBs. Extension of the assay duration to more than 7 days or using other differentiation protocols may further improve *TeraTox*'s capacity to model mesoderm and endoderm development. Omics profiling of EBs may reveal the best condition.

There are several viable options to further improve the predictivity and the explainability of the *TeraTox* model. To better distinguish between non-teratogens and teratogens, we may try to test the compounds with the *TeraTox* assay at lower concentrations (especially for non-teratogens), where the lowest concentration should be predicted to have a teratogenicity score equal to or close to zero. Multi-model data, if available, can be used to identify further relevant features beyond germ-layer genes and factors. As more and more data are collected, we may also optimize the prediction algorithm, for instance using the nearest-neighbor prediction or other variants, to benefit from the data.

Finally, the *TeraTox* assay may benefit from a better modelling of human embryo development. We may use alternative morphology-based assays of gastruloids to complement the *TeraTox* readout (96,97). Alternatively, sophisticated microphysiological systems may better mimic the maternal-placenta-embryo axis and with that may recapitulate true embryo exposure levels (98-100). In the future they may replace the 3D embryoid bodies 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 modelling of data of human embryogenesis, for instance from omics, imaging, and perturbation studies, is required to guide further optimization of the *TeraTox* assay (96,101,102).

5. Conclusion

In summary, we demonstrate that the *TeraTox* assay addresses several limitations of the industrial standard mEST assay regarding performance, species-specificity, and explainability. We believe that further optimization of the *TeraTox* assay and its routine use in drug-screening processes will lead us towards better preclinical assessment of teratogenicity.

6. Acknowledgement

We thank Kevin Michaelsen, Claudia Bossen, and Jean-Christophe Hoflack for their generous support. JDZ thanks colleagues of the Bioinformatics and Exploratory Data Analysis (BEDA) team for their input and discussions.

7. Funding

This work was supported by CEFIC, the BMBF, EFSA, and the DK-EPA (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.

8. Disclosure Statement

Some authors (MJ, JDZ, SPL, NS, PB, EK, NC and SK) are employees of F. Hoffmann-La Roche Ltd, and all authors have nothing to disclose.

9. References

1. Vargesson, N. (2015) Thalidomide-induced teratogenesis: history and mechanisms. *Birth defects research. Part C, Embryo today : reviews*, **105**, 140-156.
2. 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**.
3. ICH. (2020) ICH Harmonized Guideline Detection Of Reproductive And Developmental Toxicity For Human Pharmaceuticals S5(R3).
4. Lenz, W. and Knapp, K. (1962), *Problems of Birth Defects*. Springer, pp. 200-206.
5. Barrow, P. (2016) Revision of the ICH guideline on detection of toxicity to reproduction for medicinal products: SWOT analysis. *Reproductive Toxicology*, **64**.
6. Seiler, A.E. and Spielmann, H. (2011) The validated embryonic stem cell test to predict embryotoxicity in vitro. *Nat Protoc*, **6**, 961-978.
7. 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. *Alternatives to laboratory animals : ATLA*, **32**, 209-244.
8. Whitlow, S., Bürgin, H. and Clemann, N. (2007) The embryonic stem cell test for the early selection of pharmaceutical compounds. *ALTEX-Alternatives to animal experimentation*, **24**, 3-7.
9. Brannen, K.C., Chapin, R.E., Jacobs, A.C. and Green, M.L. (2016) Alternative Models of Developmental and Reproductive Toxicity in Pharmaceutical Risk Assessment and the 3Rs. *ILAR journal*, **57**, 144-156.

10. Brannen, K.C., Charlap, J.H. and Lewis, E.M. (2013) In Barrow, P. C. (ed.), *Teratogenicity Testing: Methods and Protocols*. Humana Press, Totowa, NJ.

11. 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.

12. 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.

13. 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.

14. 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*, e52333.

15. 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*.

16. 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.

17. 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.

18. 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.

19. 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 & molecular toxicology*, **13**, 51-66.

20. Scholz, G., Genschow, E., Pohl, I., Bremer, S., Paparella, M., Raabe, H., Southee, J. and Spielmann, H. (1999) Prevalidation of the Embryonic Stem Cell Test (EST)-A New In Vitro Embryotoxicity Test. *Toxicol In Vitro*, **13**, 675-681.

21. Scholz, G., Pohl, I., Genschow, E., Klemm, M. and Spielmann, H. (1999) Embryotoxicity screening using embryonic stem cells in vitro: correlation to in vivo teratogenicity. *Cells, tissues, organs*, **165**, 203-211.

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

23. 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. *Reproductive Toxicology*.

24. 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.

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

26. Drawnel, F.M., Zhang, J.D., Küng, E., Aoyama, N., Benmansour, F., Araujo 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 Chemical Biology*, **24**, 624-634.e623.

27. 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.

28. 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.

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

30. McGinnis, C., Schaefer, N., Kissling, T., Zumstein, T., Fattinger, C., Kolaja, K., Singer, T., Weiser, T. and Clemann, N. (2012) The Roche experience with the EST assay: Development of a new algorithm and hanging drop culture device for improved prediction and full automation. *Reproductive Toxicology*, **34**, 152.

31. 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.

32. 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. *Nature Chemical Biology*, **4**, 59-68.

33. 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. *Journal of Biomolecular Screening*, **18**, 1321-1329.

34. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.

35. 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.

36. Actelion Pharmaceuticals US, I. (2008) TRACLEER® [bosentan] *https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/021290s015lbl.pdf*.

37. Corporation, N.P. (2009) Tegretol® carbamazepine USP *https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/016608s101,018281s048lbl.pdf*.

38. Company, B.-M.S. (2016) HYDREA (hydroxyurea) HIGHLIGHTS OF PRESCRIBING INFORMATION *https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/016295Orig1s047,s048Lbl.pdf*.

39. Corporation, N.P. (2012) GLEEVEC (imatinib mesylate). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2012/021588s035lbl.pdf*.

40. Inc., R.L. (2010) ACCUTANE®(isotretinoin capsules) *https://www.accessdata.fda.gov/drugsatfda_docs/label/2010/018662s060lbl.pdf*.

41. Inc., C.P. and (2019) REDITREX (methotrexate). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/210737s000lbl.pdf*.

42. Inc., A. (2016) DEPAKENE (valproic acid). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/018081s065_018082s048lbl.pdf*.

43. Amivas. (2020) ARTESUNATE for injection, for intravenous use. *https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/213036s000lbl.pdf*.

44. GlaxoSmithKline. (2003) MYLERAN® (busulfan) *https://www.accessdata.fda.gov/drugsatfda_docs/label/2003/09386slr023_myleran_lbl.pdf*.

45. GlaxoSmithKline. (2014) TAFINLAR (dabrafenib) capsules, for oral use. *https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/202806s002lbl.pdf*.

46. Company, B.-M.S. (2010) SPRYCEL™ (dasatinib). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2010/021986s7s8lbl.pdf*.

47. Ltd, N.P. (2018) Tafinlar. *https://www.ema.europa.eu/en/documents/product-information/tafinlar-epar-product-information_de.pdf*.

48. 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.

49. 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 & development*, **16**, 2743-2748.

50. Evans, T.J. (2007) In Gupta, R. C. (ed.), *Veterinary Toxicology*. Academic Press, Oxford, pp. 206-244.

51. 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.

52. Kameoka, S., Babiarz, J., Kolaja, K. and Chiao, E. (2013) A High-Throughput Screen for Teratogens Using Human Pluripotent Stem Cells. *Toxicological Sciences*, **137**, 76-90.

53. 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. *Journal of medicinal chemistry*, **56**, 2700-2704.

54. 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. *Journal of Anatomy*, **230**, 701-709.

55. 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.

56. 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. *British journal of clinical pharmacology*, **85**, 2856-2863.

57. FDA. (2019) Amoxicillin Use by Pregnant and Lactating Women Exposed to Anthrax. *https://www.fda.gov/drugs/bioterrorism-and-drug-preparedness/amoxicillin-use-pregnant-and-lactating-women-exposed-anthrax*.

58. 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. *British journal of clinical pharmacology*, **83**, 2557-2571.

59. McGuff Pharmaceuticals, I. (2017) ASCOR® (Acsorbic Acid). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/209112s000lbl.pdf*.

60. Rumbold, A., Ota, E., Nagata, C., Shahrook, S. and Crowther, C.A. (2015) Vitamin C supplementation in pregnancy. *Cochrane Database of Systematic Reviews*.

61. 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. *Journal of obstetrics and gynaecology : the journal of the Institute of Obstetrics and Gynaecology*, **34**, 392-399.

62. Inc, D.o.P. (2002) ZYRTEC®(cetirizine hydrochloride). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2002/19835s15,%2020346s8lbl.pdf*.

63. Inc., T.P.U. (2014) cyproheptadine hydrochloride tablets usp. *https://www.iodine.com/drug/periactin/fda-package-insert*.

64. USA, T.P. (2007) CYPROHEPTADINE HYDROCHLORIDE. *https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/087056s045lbl.pdf*.

65. Merck & Co., I. (2019) DECADRON® (DEXAMETHASONE TABLETS, USP) *https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/011664s064lbl.pdf*.

66. Ltd, M.P.I.P. (2008) DORYX® (doxycycline hyclate). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2008/050795s005lbl.pdf*.

67. Inc, P. (2007) Motrin®Ibuprofen Tablets, USP *https://www.accessdata.fda.gov/drugsatfda_docs/label/2007/017463s105lbl.pdf*.

68. Adams, S.S., Bough, R.G., Cliffe, E.E., Lessel, B. and Mills, R.F.N. (1969) Absorption, distribution and toxicity of ibuprofen. *Toxicology and Applied Pharmacology*, **15**, 310-330.

69. Company, B.-M.S. (2017) GLUCOPHAGE® (metformin hydrochloride) Tablets. *https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/020357s037s039,021202s021s023lbl.pdf*.

70. King Pharmaceuticals, I. (2012) Bicillin® C-R (penicillin G benzathine and penicillin G procaine injectable suspension) *https://www.accessdata.fda.gov/drugsatfda_docs/label/2012/050138s234lbl.pdf*.

71. Dashe, J.S. and Gilstrap, L.C., 3rd. (1997) Antibiotic use in pregnancy. *Obstetrics and gynecology clinics of North America*, **24**, 617-629.

72. Inc., F.P. (2008) Endometrin® (progesterone). *https://www.accessdata.fda.gov/drugsatfda_docs/label/2008/022057s001lbl.pdf*.

73. Sergejew, T. (2015) Evaluation of a human embryonic stem cell-based screening assay for the identification of teratogenic pharmaceuticals. *THESIS FOR MASTER OF ADVANCED STUDIES IN TOXICOLOGY*.

74. Scutari, M. (2010) Learning Bayesian Networks with the bnlearn R Package. *2010*, **35**, 22.

75. 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 Journal*, **9**, 378-400.

76. Lüdecke, D. (2018) ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *The Journal of Open Source Software*, **3**.

77. Fabrigar, L.R. and Wegener, D.T. (2012) *Exploratory factor analysis*. Oxford University Press, Oxford.

78. Hochreiter, S., Clevert, D.-A. and Obermayer, K. (2006) A new summarization method for affymetrix probe level data. *Bioinformatics*, **22**, 943-949.

79. BUTTE, A.J. and KOHANE, I.S., *Biocomputing 2000*, pp. 418-429.

80. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

81. Khanin, R. and Wit, E. (2006) How Scale-Free Are Biological Networks. *Journal of Computational Biology*, **13**, 810-818.

82. Broido, A.D. and Clauset, A. (2019) Scale-free networks are rare. *Nature Communications*, **10**, 1017.

83. 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. *Clinical Pharmacology & Therapeutics*, **107**, 871-885.

84. 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. *Information Fusion*, **58**, 82-115.

85. Waldmann, T., Grinberg, M., König, A., Rempel, E., Schildknecht, S., Henry, M., Holzer, A.-K., Dreser, N., Shinde, V., Sachinidis, A. *et al.* (2016) Stem Cell Transcriptome Responses and Corresponding Biomarkers That Indicate the Transition from Adaptive Responses to Cytotoxicity. *Chemical Research in Toxicology*, **30**, 905-922.

86. 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.

87. 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.

88. 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.

89. 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.

90. 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.

91. 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**.

92. 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.

93. 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.

94. 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.

95. 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 research*, **48**, 12577-12592.

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

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

98. 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)*, e2100609.

99. 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.

100. 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**.

101. 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.

102. 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. *Archives of Toxicology*.