Environment International 158 (2022) 106947

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Quantitative high-throughput phenotypic screening for environmental estrogens using the E-Morph Screening Assay in combination with *in silico* predictions

Saskia Klutzny [a], Marja Kornhuber [a,b], Andrea Morger [c], Gilbert Schönfelder [a,d], Andrea Volkamer [c], Michael Oelgeschläger [a], Sebastian Dunst [a,*]

[a] *Experimental Toxicology and ZEBET, German Federal Institute for Risk Assessment (BfR), German Centre for the Protection of Laboratory Animals (Bf3R), Berlin, Germany*
[b] *Freie Universität Berlin, Berlin, Germany*
[c] *In silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany*
[d] *Institute of Clinical Pharmacology and Toxicology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany*

ARTICLE INFO

Handling Editor: Olga Kalantzi

ABSTRACT

Background: Exposure to environmental chemicals that interfere with normal estrogen function can lead to adverse health effects, including cancer. High-throughput screening (HTS) approaches facilitate the efficient identification and characterization of such substances.
Objectives: We recently described the development of the E-Morph Assay, which measures changes at adherens junctions as a clinically-relevant phenotypic readout for estrogen receptor (ER) alpha signaling activity. Here, we describe its further development and application for automated robotic HTS.
Methods: Using the advanced E-Morph Screening Assay, we screened a substance library comprising 430 toxicologically-relevant industrial chemicals, biocides, and plant protection products to identify novel substances with estrogenic activities. Based on the primary screening data and the publicly available ToxCast dataset, we performed an *in silico* similarity search to identify further substances with potential estrogenic activity for follow-up hit expansion screening, and built seven *in silico* ER models using the conformal prediction (CP) framework to evaluate the HTS results.
Results: The primary and hit confirmation screens identified 27 'known' estrogenic substances with potencies correlating very well with the published ToxCast ER Agonist Score ($r = +0.95$). We additionally detected potential 'novel' estrogenic activities for 10 primary hit substances and for another nine out of 20 structurally similar substances from *in silico* predictions and follow-up hit expansion screening. The concordance of the E-Morph Screening Assay with the ToxCast ER reference data and the generated CP ER models was 71% and 73%, respectively, with a high predictivity for ER active substances of up to 87%, which is particularly important for regulatory purposes.
Discussion: These data provide a proof-of-concept for the combination of *in vitro* HTS approaches with *in silico* methods (similarity search, CP models) for efficient analysis of large substance libraries in order to prioritize substances with potential estrogenic activity for subsequent testing against higher tier human endpoints.

1. Introduction

Endocrine-disrupting chemicals (EDCs) are a group of exogenous substances that interfere with the endocrine system, leading to adverse health effects, including cancer (WHO/IPCS, 2002). Global cancer burden has substantially increased over the last decades and incidence

* Corresponding author at: German Federal Institute for Risk Assessment (BfR), German Centre for the Protection of Laboratory Animals (Bf3R), Max-Dohrn-Straße 8-10, 10589 Berlin, Germany
E-mail address: Sebastian.Dunst@bfr.bund.de (S. Dunst).

https://doi.org/10.1016/j.envint.2021.106947
Received 14 July 2021; Received in revised form 14 October 2021; Accepted 18 October 2021
Available online 28 October 2021

rates are projected to further rise in the future (Sung et al. 2021; Wild et al. 2020). In 2020, female breast cancer has been the most commonly diagnosed cancer worldwide (Sung et al. 2021) and estrogens are an important risk factor (Yager and Davidson 2006). Thus, the identification of exogenous substances that mimic estrogen function and subsequent reduction of human exposure to such substances are important measures for the effective prevention of endocrine-related cancers such as breast cancer.

Environmental sources of substances with estrogenic activity are manifold and include consumer products and food packaging materials, preservatives, food additives, and pesticides, but are also naturally found in food (e.g. phytoestrogens) (Paterni et al. 2017). In regulatory toxicology, the detection of potential substance-related adverse health effects still calls for traditional *in vivo* test guideline studies that mainly use rodents to evaluate the potential hazards of single substances (OECD 2001; 2018a; b; c). However, such cost- and time-consuming *in vivo* studies not necessarily mimic human-relevant physiological and disease conditions. In addition, the ethical issues related to animal testing in general and the high numbers of test animals needed further fuel the rising need and interest in human-relevant alternative *in silico*, *in chemico*, and *in vitro* test methods to reduce and eventually replace animal testing according to the 3Rs principle (Russell and Burch 1959).

Following the assumption that structurally similar substances can have similar toxicological effects (Maggiora et al. 2014), risk assessment commonly uses read-across approaches to effectively reduce costs and animal testing (Carrio et al. 2016; Hemmerich and Ecker 2020; Raies and Bajic 2016). Indeed, various established QSAR models, which represent statistical *in silico* models that relate a set of structural descriptors of a substance to its biological activity, can support read-across (Ma et al. 2015; Tropsha 2010). Still, the prediction of toxicity from structural similarity remains challenging for risk assessment because *in silico* predictions alone do not yet sufficiently fulfill information requirements for complex human health endpoints.

In recent years, machine learning approaches gained momentum, which use the increasingly comprehensive *in chemico* and *in vitro* test data in an iterative process with growing certainty to identify combinations of substance features that may lead to a specific toxicological effect (Gayvert et al., 2016; Huang and Xia, 2017; Mayr et al., 2016). A special case of machine learning is conformal prediction (CP), which adds confidence estimation to the model predictions (Alvarsson et al. 2021; Norinder et al. 2014; Vovk et al. 2005). The CP framework is built on top of a machine learning method and uses an additional calibration step based on experimental test results to determine the confidence when making predictions on new data. CP models have recently been intensively and successfully built and applied to toxicological questions (Morger et al. 2020; Morger et al. 2021; Norinder et al. 2016; Svensson et al. 2017b; Zhang et al. 2021).

The increasing quantity and diversity of chemicals that are produced and marketed worldwide stimulate the establishment of high-throughput screening (HTS) research programs that use *in chemico* and *in vitro* assays to efficiently generate comprehensive concentration–response information for a large number of substances. For example, the U.S. EPA Toxicity Forecaster (ToxCast) project generated screening data for over 10,000 environmental chemicals that were tested in hundreds of HTS assays addressing toxicological and endocrine endpoints in order to rank and prioritize substances for subsequent *in vivo* testing (Dix et al. 2007; Judson et al. 2010; Reif et al. 2010; Rotroff et al. 2013). These screening data have further been integrated into an *in silico* ToxCast ER pathway model, which converts results from 18 automated ER screening assays into a relative ER bioactivity score ranging from 0.00 (no activity) to 1.00 (bioactivity of the reference substance 17-alpha-Ethinylestradiol) (Browne et al. 2015; Judson et al. 2015). More recently, Judson *et al.* demonstrated that a reduced set of four out of the originally 18 ER screening assays achieves a comparable performance (Judson et al. 2017). Furthermore, the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) developed another *in silico* consensus model for prediction of ER binding, agonistic, and antagonistic activities of chemicals (Mansouri et al. 2016). This CERAPP ER consensus model integrates 48 individual computational models using different QSAR and structure-based approaches, which have been trained and optimized using the relative potency information from the ToxCast ER Agonist Model (Browne et al. 2015; Judson et al. 2015).

The *in chemico* and *in vitro* HTS assays that provide the data for the ToxCast ER pathway model each cover single, mechanistic events of estrogen signaling (ER binding, ER dimerization, regulation of gene expression, and cell proliferation) but the immediate relevance of the derived data regarding adverse effects, including human cancer, is still limited. Hence, development and application of novel cell-based test methods that combine HTS capability with more human-relevant, functional endpoints can support a more direct extrapolation of the test results to the complex signaling events and regulatory mechanisms that drive adverse effects including cancer progression and metastasis. We have recently shown that the cell-based E-Morph Assay provides such an endpoint, i.e., the machine learning-based analysis of estrogen-dependent phenotypic changes at adherens junctions (AJ) (Bischoff et al. 2020; Kornhuber et al. 2021). The E-Morph Assay is based on the observation that the inhibition of ER signaling in an MCF-7 breast cancer cell line led to a prominent reorganization of AJs and induced the clustering of the AJ protein E-cadherin (E-Cad), which could be reverted by co-treatment with estrogenic substances (Bischoff et al. 2020). These changes in cell morphology correlated with increased cellular stiffness and decreased cell motility, with deregulation of these two parameters often being associated with breast cancer progression and metastasis (Bischoff et al. 2020). In addition, we could describe comparable changes in E-Cad localization in clinical breast cancer tissue samples supporting the clinical relevance of the assay endpoint.

In the present study, we optimized the E-Morph Assay for automated robotic HTS and used this advanced E-Morph Screening Assay to analyze a substance library comprising 430 toxicologically-relevant industrial chemicals, biocides and plant protection products that are reported to act through various nuclear hormone receptors. Using our HTS data in combination with already publicly available *in chemico* and *in vitro* ToxCast data as well as *in silico* prediction approaches using the CP framework, we could further identify additional, novel substances with potential estrogenic activity.

2. Materials and methods

2.1. Cell line and cell culture conditions

The MCF-7/E-Cad-GFP cell line (de Beco et al. 2009; 2020) that stably expresses a fluorescent E-Cadherin-GFP fusion protein was kindly provided by Sylvie Coscoy (Laboratoire Physico-Chimie Curie, Institut Curie, PSL Research University - Sorbonne Universités, UPMC-CNRS, Paris, France).

Routine cell cultures were maintained at 37 °C with 5% CO_2 in normal-serum medium containing Dulbecco's modified Eagle's medium (DMEM, low glucose, pyruvate, no glutamine, no phenol red) (Gibco/Thermo Fisher Scientific, Waltham, MA, USA), 10% (v/v) Fetal Bovine Serum (FBS, S0615, Estradiol levels: 22.3 pg/ml) (Biochrom/Merck, Darmstadt, Germany), 2 mM stable glutamine (Gibco/Thermo Fisher Scientific), 100 μg/ml streptomycin / 100 U/ml penicillin (Biochrom/Merck), and 0.4 mg/ml geneticin (Gibco/Thermo Fisher Scientific). Cells were sub-cultured over a maximum of 10–12 passages, and regularly tested using the Eurofins Genomics mycoplasma test service (Eurofins Genomics, Ebersberg, Germany).

Experiments were performed in reduced-serum medium as described above but containing only 5% (v/v) FBS to minimize background estrogen levels and potential test chemical binding to serum lipids and proteins in the exposure medium. The final estradiol concentration in reduced-serum medium (4.1 pM) was in the range of physiological

serum levels of postmenopausal women (Rothman et al. 2011). If not otherwise stated, cells were seeded into multi-well plates at suitable concentrations to achieve 80–90% confluency after 24 h. Cells were then exposed to reduced-serum medium containing the respective test chemical in combination with the anti-estrogen Fulvestrant (Fulv, 10 nM) (Sigma-Aldrich/Merck, Darmstadt, Germany) for 48 h, followed by the sample preparation procedure. Experimental controls included the solvent control, the Fulv control containing 10 nM Fulv only, and the co-treatment (reactivity) control containing 10 nM Fulv + 10 μM Estrone (Sigma-Aldrich/Merck) (each in reduced-serum medium). In all experiments, the solvent control corresponds to the respective experimental conditions, excluding Fulv and test chemicals. The DMSO (Sigma-Aldrich/Merck) concentration in the solvent control was always adjusted to the highest DMSO concentration used in the experiment, i.e. in the range of 0.1–0.4% depending on the experimental setting.

2.2. Quantitative PCR

Cells were seeded into 12-well plates at a concentration of $4x10^5$ cells/well in 1 ml reduced-serum medium and exposed to test substances as described above. RNA extraction (RNeasy Kit, Qiagen, Hilden, Germany), cDNA synthesis (High-Capacity cDNA Reverse Transcription Kit, Applied Biosystems/Thermo Fisher Scientific, Waltham, MA, USA), and quantitative PCR (qPCR) (PowerUp SYBR Green Master Mix, Applied Biosystems/Thermo Fisher Scientific) were conducted according to the manufacturers protocols using a QuantStudio 7 Flex Real-Time PCR System (Applied Biosystems/Thermo Fisher Scientific) (40 cycles; denaturation for 15 s at 95 °C; annealing, extension, and fluorescence read for 1 min at 60 °C). RNA concentrations (A260) and purity ratios (A260/A280 and A260/A230) were determined using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific). Purity ratios of ~ 2.0 (A260/A280) and 2.0–2.2 (A260/A230) were generally considered acceptable. cDNA synthesis was performed using 1 μg RNA and RT random primers (High-Capacity cDNA Reverse Transcription Kit, Applied Biosystems/Thermo Fisher Scientific). For qPCR, 1 μl of 1:10 diluted (water) cDNA was added to 10 μl master mix containing water, primers and SYBR Green. RNA expression levels (fold change) were calculated according to the $\Delta\Delta C_T$ method (Livak and Schmittgen 2001). Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein, Zeta (YWHAZ) was used as housekeeping gene. If not otherwise stated, each experiment was performed in technical triplicates and in at least three independent repetitions. Primers used (5′-3′ orientation):

BCL2L1 (CAGCTTGGATGGCCACTTAC, TGCTGCATTGTTCCCATAGA);

TFF1 (CATCGACGTCCCTCCAGAAGAG, CTCTGGGACTAATCACCGTGCTG);

PGR (TCAACTACCTGAGGCCGGAT, GCTCCCACAGGTAAGGACAC);

AREG (TGGATTGGACCTCAATGACA, TAGCCAGGTATTTGTGGTTCG);

ESR1 (CCACCAACCAGTGCACCATT, GGTCTTTTCGTATCCCACCTTTC);

GFP (AAGCTGACCCTGAAGTTCATCTGC, CTTGTAGTTGCCGTCGTCCTTGAA);

CDH1 (AGGAGCCAGACACATTTATGGAA, GCTGTGTACGTGCTGTTCTTCAC);

mCdh1 (AACCCAAGCACGTATCAGGG, GAGTGTTGGGGGCATCATCA);

YWHAZ (ACTTTTGGTACATTGTGGCTTCAA, CCGCCAGGACAAACCAGTAT).

2.3. Western blot

Cells were seeded into 6 or 12-well plates at a concentration of $1x10^6$ or $4x10^5$ cells/well in 2 ml or 1 ml reduced-serum medium and exposed to substances as described above. For protein extraction, cells were washed with ice-cold Phosphate-buffered saline (PBS) and scraped in 100–200 μl lysis buffer (50 mM Tris/HCl pH 7.4, 150 mM NaCl, 0.1% (w/v) Na-deoxycholate, 0.1% (w/v) SDS, 1% (v/v) IGEPAL CA-630/NP-40, 5 mM EDTA pH 8.0, 5 mM EGTA, 1X cOmplete Protease Inhibitor Cocktail (Roche, Basel, Switzerland), PhosSTOP Phosphatase Inhibitor Cocktail (Roche) and incubated for 30 min on ice. Lysates were centrifuged at 13,000 g and 4 °C for 10 min, and the supernatant was collected. Total protein concentrations were determined using a Pierce BCA Protein Assay Kit (Thermo Scientific/Thermo Fisher Scientific, Waltham, MA, USA) and a BSA standard (Thermo Fisher Scientific) according to the manufacturer's instructions. Protein lysates were separated by SDS-PAGE using Mini-PROTEAN precast gels (4–15% polyacrylamide) (Bio-Rad Laboratories, Hercules, CA, USA) according to the manufacturer's instructions. Proteins were transferred onto nitrocellulose membranes (Bio-Rad Laboratories) using a semi-dry Trans-Blot Turbo Transfer System (1.3 A per gel, 25 V for 7 min) (Bio-Rad Laboratories). Membranes were blocked with 5% low-fat milk powder for 60 min, rinsed in Tris-buffered saline containing Tween 20 (TBS-T) (TBS, 0.1% Tween 20), and incubated with primary and secondary antibodies in 0.6% low-fat milk powder in TBS-T (TBS, 0.1% Tween 20) over night at 4 °C and for 3 h at room temperature, respectively. Antibodies/dyes used: mouse anti-E-Cad (1:1,000) (Clone 36, BD Biosciences, Franklin Lakes, NJ, USA) and HRP-conjugated goat anti-mouse secondary antibody (1:10,000) (Jackson ImmunoResearch, West Grove, PA, USA). Protein detection was carried out using a Pierce ECL Western Blotting Substrate (Thermo Scientific/Thermo Fisher Scientific) in a Fusion Solo S (VWR, Radnor, PA, USA) imaging system. Coomassie Brilliant Blue (Bio-Rad Laboratories) total protein staining of nitrocellulose membranes was used as loading control (Welinder and Ekblad 2011). Semi-quantitative densitometric analysis of western blot bands was performed using the FIJI software (Schindelin et al. 2012). The band intensities were normalized to the respective Coomassie total protein staining of each lane. The results from each treatment condition were then normalized to the solvent control.

2.4. siRNA knockdown

Cells were seeded into 6 or 12-well plates at a concentration of $1x10^6$ or $4x10^5$ cells/well in 2 ml or 1 ml reduced-serum medium and exposed to substances as described above. Transfections were carried out using the HiPerFect Transfection Reagent (Qiagen) and a mix of four *ESR1* siRNAs (FlexiTube GeneSolution GS2099, Quiagen) (10 nM) with different target sequences (SI02781401; SI03114979; SI03065615; SI00002527). Cells were transfected at the time of cell seeding according to the manufacturer's reverse-transfection protocol.

2.5. ER binding experiments

Cells were seeded into 96-well plates and exposed to substances as described above. Transfections were carried out using the FuGENE HD Transfection Reagent (Promega, Madison, WI, USA), the pBIND-ERα [*hRluc*] vector (50 ng), and the pGL4.35[*luc2P/9XGAL4 UAS/Hygro*] vector (50 ng) (both Promega) according to the manufacturer's protocol. The binding of a test substance with estrogenic activity to a fusion protein containing an estrogen receptor-ligand binding domain (ER-LBD) and a yeast Gal4 DNA-binding domain (Gal4-DBD) (pBIND-ERα [*hRluc*] vector) led to the expression of an UAS-controlled Firefly luciferase reporter protein (pGL4.35[luc2P/9XGAL4UAS/Hygro] vector), which was detected using the Dual-Glo Luciferase Reagent (Promega). The detected Firefly luminescence (pGL4.35[luc2P/9XGAL4UAS/Hygro] vector) was normalized to the Renilla luminescence (pBIND-ERα [*hRluc*] vector) to derive a relative signal intensity.

2.6. E-Morph Screen: Cell seeding and test substance exposure scenarios

Cells were seeded into CellCarrier-96 Ultra Microplates (PerkinElmer, Waltham, MA, USA) at a concentration of $9x10^4$ cells/well in

225 µl reduced-serum medium, grown until 80–90% confluency for 24 h, and then exposed to 250 µl reduced-serum medium containing each test chemical in combination with 10 nM Fulv for 48 h.

All 430 test substances (Sigma-Aldrich/Merck) of the BfR-ChemLibrary were previously dissolved in DMSO (Sigma-Aldrich/Merck) at a stock concentration of 10 mM and stored at the *Compound Management Unit* of the Leibniz Institute of Molecular Pharmacology (FMP, Berlin, Germany). For this project, a copy of the BfR-ChemLibrary was provided by the FMP on five 96-well microplates (Greiner Bio-One, Frickenhausen, Germany) along with an empty column for the assay controls. The preparation of the exposure medium and its application to cells was performed using a JANUS Automated Liquid Handling Workstation (PerkinElmer) and customized treatment protocols written in WinPREP (PerkinElmer).

For the hit selection (primary) screen, 2 µl of the test substance (10 mM) were transferred into an empty 96-well microplate (Greiner Bio-One) and then dissolved (1:100) in 198 µl reduced-serum medium containing 100 nM Fulv. Subsequently, 25 µl of the exposure medium containing the diluted test substance (100 µM) were then transferred from the compound plates to the cell culture assay plates containing 225 µl reduced-serum medium to achieve a final test substance concentration of 10 µM and a final Fulv concentration of 10 nM. Considering that the nominal concentration of a test chemical does not necessarily reflect the concentration at the target site due to potential partitioning of test chemicals to other extracellular compartments in *in vitro* assays (Proença et al., 2021), a starting concentration of 10 µM is often used for hit selection in comparable HC/HT screening projects in order to maximize exposure of cells to the test substance. Sufficiently high exposure levels ensure confidence in negative test results and are particularly important for detection of substances with weak estrogenic activities, such as industrial chemicals, that were, in contrast to pharmaceuticals, not designed to act on the ER pathway.

For the hit confirmation (potency) and the hit expansion screens, 9 µl of the test substance (10 mM) were transferred into an empty 96-well microplate (Greiner Bio-One) and then dissolved (1:33) in 291 µl reduced-serum medium containing 100 nM Fulv. From this start concentration (300 µM), serial dilutions were generated at a 1:3 ratio in reduced-serum medium containing 100 nM Fulv. Subsequently, 25 µl of the exposure medium containing the diluted test substance (300 µM to 10 pM) was then transferred from the compound plates to the cell culture assay plates containing 225 µl reduced-serum medium to achieve a final test substance concentration of 30 µM to 1 pM and a final Fulv concentration of 10 nM.

For all screening approaches, the three solvent control wells of each plate contained 0.2% (primary screen) or 0.4% (hit confirmation screen) DMSO, the three Fulv control wells contained 10 nM Fulv, and the two co-treatment (reactivity) control wells contained 10 nM Fulv + 10 µM Estrone (each in reduced-serum medium).

2.7. E-Morph Screen: Fluorescence microscopy and quantitative image analysis

The preparation of the cells for fluorescence microscopy was performed using a JANUS Automated Liquid Handling Workstation (PerkinElmer) and an ELx405 Select CW Microplate Washer (BioTek Instruments, Winooski, VT, USA). After treatment for 48 h, the cells were stained in PBS containing 1 µM CellTrace Far Red (Molecular Probes/Thermo Fisher Scientific, Waltham, MA, USA) to visualize the cell–cell contact morphology according to (Kornhuber et al. 2021) as an internal quality control and 2 µg/ml Hoechst 33,342 (Molecular Probes/Thermo Fisher Scientific) to label nuclei for 20 min at 37 °C with 5% CO_2, then washed twice with PBS, fixed with 4% formaldehyde solution for 15 min at room temperature, and finally washed again with PBS. During this procedure, the E-Cad-GFP signal was preserved and did not require additional staining.

Cells were subsequently imaged with an Opera Phenix High-Content Screening System (PerkinElmer) using a 20x air objective (NA 0.4) at three standardized positions per well and three or four optical sections with 4 µm or 3 µm spacing per position. Image analysis was performed using the integrated Harmony software (PerkinElmer) and customized image analysis routines (Fig. S1A). First, nuclei were identified using the Hoechst 33,342 channel to define each cell. Nuclei touching the edge of the image were excluded from further analysis. Next, cell outlines were identified using the GFP channel and the E-Cad-GFP signal intensity was measured for each cell. Finally, a mean E-Cad-GFP signal intensity was calculated across all cells per well.

For visualization of concentration-response curves of relative E-Cad-GFP signal intensities ($SI^{E\text{-}Cad\text{-}GFP}$), the mean E-Cad-GFP signal intensity (SI) per well (SI_{well}) was normalized to the average SI (SI_{avr}) of the corresponding three solvent control wells on each plate according to (Malo et al. 2006):

$$SI^{E\text{-}Cad\text{-}GFP} = \frac{SI_{well}^{Substance}}{SI_{avr}^{Solv}} * 100$$

2.8. E-Morph Screen: Automated data evaluation

The process automation software KNIME [v4.1.2] (Berthold et al. 2008) was used to build a customized pipeline (Fig. S1B; File S1) for fast and efficient automated processing, evaluation, and statistical analysis of the quantitative image data obtained from the individual screens. Briefly, this KNIME workflow retrieved all *.txt* files that were exported from the Harmony software (PerkinElmer) into a specified folder and executed the following steps in a loop function: a) import *.txt* files and convert to tables b) adjust table columns and rows (e.g. remove unnecessary columns), c) merge all measurement tables into a global table, and d) join measurement data with the plate assignment metadata (e.g. substance name and concentration).

In order to detect potential cytotoxic substance effects on cell viability (CV), the mean number of nuclei (N) per well (N_{well}) was normalized to the average N (N_{avr}) of the corresponding three Fulv control wells on each plate according to (Malo et al. 2006):

$$CV = \frac{N_{well}^{Substance}}{N_{avr}^{Fulv}} * 100\%$$

Substances leading to a CV < 75% (i.e., representing a ≥ 25% reduction of the number of nuclei compared to the 10 nM Fulv control) in at least two out of three runs were assigned to the group of 'Toxic substances'. For substances leading to a CV ≥ 75%, the mean E-Cad-GFP signal intensities were further analyzed.

To identify potential estrogenic substances, the mean E-Cad-GFP signal intensity (SI) per well (SI_{well}) was normalized to both the average SI (SI_{avr}) of the corresponding three solvent control wells (SI = 100) AND the three Fulv control wells (SI = 0) on each plate according to (Malo et al. 2006):

$$SI = \frac{SI_{avr}^{Fulv} - SI_{well}^{Substance}}{SI_{avr}^{Fulv} - SI_{avr}^{Solv}} * 100$$

Substances leading to an SI ≥ 20 in at least two out of three runs were considered as potential estrogenic substances in the primary screen. The image data of these substances were furthermore visually assessed for ambiguous results and potential imaging artifacts.

For quality assessment, the signal separation (effect size) between the solvent control and Fulv control and the deviation of values within each control group was determined for each plate and run. The Z'-factor (Z') was calculated based on the average SI (SI_{avr}) and the standard deviation (SI_{sd}) of both the three Fulv control and the three solvent control wells on each plate according to (Iversen et al. 2006; Zhang et al. 1999):

$$Z' = \frac{\left(SI_{avr}^{Solvent} - 3SI_{sd}^{Solvent}\right) - \left(SI_{avr}^{Fulv} + 3SI_{sd}^{Fulv}\right)}{SI_{avr}^{Solvent} - SI_{avr}^{Fulv}} = 1 - \frac{3\left(SI_{sd}^{Solvent} + SI_{sd}^{Fulv}\right)}{SI_{avr}^{Solvent} - SI_{avr}^{Fulv}}$$

The acceptance criterion for a valid run was Z' > 0.5.

2.9. Data visualization and statistical analyses

All quantitative data were exported into Excel (Microsoft, Redmond, WA, USA)-readable files. Graphical visualizations and statistical analyses of data were performed using Prism 8 (GraphPad Software, San Diego, CA, USA). Quantitative data were plotted using descriptive statistical indexes, i.e. mean and standard deviation. SI concentration–response curves from the hit confirmation and hit expansion screens were fitted using the non-linear fit algorithm (four parameters, variable hill slope) to calculate half-maximal concentrations (EC50). The Pearson correlation coefficient (r) has been used to measure the strength of association and the direction of the relationship between the determined EC50 values of the substances tested in the E-Morph Assay and the ToxCast ER Agonist Score. Statistical methods used for computational tools are described below in detail. Figures were generated using Illustrator CC 2020 (Adobe, San Jose, CA, USA).

2.10. Computational predictions

The *in silico* toxicity prediction pipeline follows the main steps from the previously published KnowTox project (Morger et al. 2020). In this study, the KnowTox pipeline was slightly adapted for the identification of substances with estrogenic activities and the individual steps are briefly explained in the following. A more detailed description of the underlying concepts is available in the original publication (Morger et al. 2020).

2.10.1. Dataset and preprocessing

2.10.1.1. ToxCast dataset. The publicly available U.S. EPA ToxCast dataset, comprising 8,390 chemicals tested against up to 1,092 endpoints, was downloaded from U.S. EPA's Center for Computational Toxicology and Exposure (U.S. EPA, 2017) and used for training of the *in silico* conformal prediction (CP) models and for the similarity search. In addition to the data preparation steps described in (Morger et al. 2020), canonical SMILES were extracted from the PubChem database using the PubChem PUG REST API (Kim et al. 2015). If no canonical SMILES were available from PubChem, the original ToxCast SMILES were retained.

2.10.1.2. Standardization. First, all instances (molecules and mixtures comprising multiple chemicals) were standardized using the IMI eTOX project standardizer tool (Atkinson 2014) applying the following steps: discard non-organic compounds, neutralize, apply certain structure standardization rules (e.g. handling of tautomers, shifting protons between heteroatoms), neutralize, and remove (mainly organic) salts. Second, standardized molecules and mixture components with less than four heavy atoms, as well as all remaining mixtures, were discarded. This standardized dataset, containing 7,911 molecules, was used as a basis for the similarity search. For subsequent read-across support, measured activities of these molecules in seven ToxCast screening assays covering relevant ER-related endpoints (see Table 4) were considered.

To train CP models on seven estrogen receptor assay datasets, the measured activities (binary) in these assays were assigned to the standardized molecules. Next, each molecule was represented as InChI (International Chemical Identifier), a standardized format, to recognize molecules that appear more than once in the dataset. Chemicals with duplicate InChIs were merged and measurements were aggregated by median (median = 0.5 was discarded). This resulted in a data frame of 7,135 compounds tested on up to seven endpoints.

2.10.1.3. Descriptor calculation. As input for similarity search and CP, descriptors were calculated for all molecules using the RDKit Python library [v.2020.03.1] (Landrum 2006). For similarity search, the circular-environment based Morgan fingerprint (1024 bits, radius 3) and the SMARTS-pattern based MACCS keys (167 bits) were calculated and concatenated. For CP, the above fingerprints were further extended with 200 physicochemical descriptors calculated using the RDKit library. The physicochemical descriptors were normalized based on mean and standard deviation of the physicochemical descriptors of substances for which ToxCast assay data was available. Further molecular descriptors used for the analysis of bisphenols (i.e. MorganCount, MACCS, pharmacophore fingerprints) were also calculated using the RDKit functionalities.

2.10.2. In silico methods

2.10.2.1. Similarity search and read-across support. To support read-across with *in vitro* activity information from similar molecules, a similarity search was implemented using RDKit functionalities. For a query molecule, the Tanimoto similarity to all molecules in the standardized ToxCast data set was calculated, based on the above-described descriptors. The ten most similar compounds were returned together with their Tanimoto similarity and the maximum common substructure with the query molecule.

2.10.2.2. Conformal prediction (CP). CP is a framework on top of a machine learning algorithm, which has the advantage to provide a measure of confidence to the prediction as it includes an additional calibration step (Alvarsson et al. 2021; Norinder et al. 2014; Vovk et al. 2005). Therefore, besides the proper training set, an additional calibration set is needed. By comparing the predictions for a query compound with the predictions already made for the calibration set, the algorithm calculates how well the new prediction conforms to the pre-calculated data points per class (i.e., binary: active and inactive, mondrian classification (Sun et al. 2017)) using calculated p-values.

Given that the training and test data are exchangeable, conformal predictors are designed to conform to a pre-defined maximum error rate. This error rate (significance level) functions as a threshold, so the CP output is a prediction set, which contains all classes for which the p-value is higher than the significance level. For a binary classification problem with classes '0' and '1', the possible prediction outputs are: 'single class' ({0},{1}), 'both class' ({0,1}), or an empty prediction set ({}). A more detailed description of CP, and specifically the use of an additional normalizer model to improve the applicability of the CP models to unseen data, is provided in (Morger et al. 2020).

For each of the ER related endpoints, CP models (nonconformist Python library (Linusson 2015)) were trained and evaluated within a fivefold cross-validation. Thus, the ToxCast data was randomly and stratified split into five parts. In each fold, 80% training and 20% test data were used and an aggregated conformal predictor (ACP) (Carlsson et al. 2014) with 20 loops was initialized. In every ACP loop, the training data was further split into a proper training (70%) and a calibration set (30%) and a random forest model (500 estimators, else default parameters, scikit-learn Python library [v.0.22.2] (Pedregosa et al. 2011)) was trained on the proper training set. The predictions were calibrated using the calibration set, inverse probability error function, and mondrian condition (Sun et al. 2017). Furthermore, the predicted values were normalized using information from the nearest neighbors of the proper set (KNNRegressor [v2.1.0] (Linusson 2015) as described in (Morger et al. 2020)). Median was used to aggregate the p-values from the 20 ACP loops, as recommended by (Linusson et al. 2017). The p-values of the cross-validation were averaged by their mean.

2.10.2.3. CP evaluation. To evaluate the CP ER models and the predictions of the hit expansion compounds, evaluation measures such as

validity, efficiency and accuracy were used. Validity was calculated as the percentage of prediction sets containing the correct class, i.e., the fraction of all 'both class' ({0,1}) and correct 'single class' ({0}, {1}) predictions. Efficiency of the models was calculated as the ratio of prediction sets only containing a single class, i.e., {0} and {1}. Accuracy was determined by the ratio of correct 'single class' classifications compared to all 'single class' predictions.

2.10.2.4. Consensus prediction. To derive a single prediction per compound over all seven CP ER models, the predictions from the individual models were merged into a so-called consensus prediction. Thus, for each prediction, the prediction set was calculated for a maximum accepted error rate of 20%. Only the efficient single class predictions were considered and the mode was calculated following a 'majority vote' principle to obtain one consensus prediction for ER agonism per substance.

2.10.2.5. Comparative docking. Docking is a structure-based modeling technique to predict the preferred orientation of a ligand when it is bound to a protein, which can be mainly divided into a placement and a scoring step (Brooijmans and Kuntz 2003). Docking was performed using the Endocrine Disruptor Monitoring tool (2019 EDMonv3) of the @TOME-2 platform, an inverse screening pipeline that was developed to study interactions between ERα and potentially ER active substances (Pons and Labesse 2009). The EDMonv3/@TOME-2 webserver provides docking of ligands into several endocrine disruption targets (collected in the database "NR_HUMAN_I90_2019M8 (Aug 2019)").

In this study, the binding modes of five bisphenol(-like) ligands (Bisphenol F, 4-Benzylphenol, 4,4''-Dihydroxybiphenyl, 4,4''-Dihydroxybenzophenone, and Bisphenol E) were analyzed when docked to ERα, for which the database contains 221 different protein structures for docking. For the docking experiment, 3D coordinates for each of the five compounds were generated using RDKit, and default parameters were used for the EDMonv3/@TOME-2 screening on ERα (H_NR3A1_ERa).

Out of 221 available structural supports for ERα, the server returned the 20 best complexes per compound. To be able to directly compare the resulting poses for the five compounds, a common target structure that was returned as one of the top 20 complexes for all queries was desired. Twelve crystal structures were returned for all five query compounds consistently. Amongst them, the 3UUA ER agonist PDB structure was chosen, which was already investigated in the analysis of bisphenol-ER interactions by (Delfosse et al. 2012). Docking results, i.e., the crystal structure of 3UUA with the best docked pose of the respective ligand, were downloaded (as *.pdb* file) from the server and further analyzed using LigandScout [v4.4.3] (Wolber and Langer 2005).

2.11. Performance calculations

For each comparison, the overall concordance of active and inactive class predictions, i.e., the proportion of all substances that are correctly classified as active ($N_{True\ Actives}$) or inactive ($N_{True\ Inactives}$) from all tested substances (N), were calculated as follows:

$$\text{Concordance} = \frac{N_{True\ Actives} + N_{True\ Inactives}}{N} * 100$$

The accuracy of active class predictions ($P_{active\ class}$), i.e., the proportion of all substances that are correctly classified as active ($N_{True\ Actives}$) from all substances that are active in the reference method ($N_{True\ Actives} + N_{False\ Inactives}$), were calculated as follows:

$$P_{active\ class} = \frac{N_{True\ Actives}}{N_{True\ Actives} + N_{False\ Inactives}} * 100$$

The accuracy of inactive class predictions ($P_{inactive\ class}$), i.e., the proportion of all substances that are correctly classified as inactive ($N_{True\ Inactives}$) from all substances that are inactive in the reference method ($N_{True\ Inactives} + N_{False\ Actives}$), were calculated as follows:

$$P_{inactive\ class} = \frac{N_{True\ Inactives}}{N_{True\ Inactives} + N_{False\ Actives}} * 100$$

3. Results and Discussion

3.1. E-Cadherin-GFP cell membrane signal intensity as a novel readout to efficiently measure estrogen signaling activity

As described in (Kornhuber et al. 2021), the E-Morph Assay allows the identification and characterization of estrogenic substances based on quantitative changes in the morphology of cell–cell contacts at the level of AJs in the MCF-7/vBOS breast cancer cell line, which occur 24–48 h after exposure to an anti-estrogenic compound such as Fulvestrant (Fulv). In this first description of the assay, the cell–cell contact morphology was visualized using live-cell staining and analyzed by applying a quantitative image analysis pipeline with an integrated classification model (Kornhuber et al. 2021).

In order to improve the HTS capability of the E-Morph Assay and to streamline the visualization and analysis procedures, we now selected a MCF-7 cell line that stably expressed an *E-Cad-GFP* transgene encoding for the mouse E-Cad fused to GFP, which has been shown to actively engage in the AJ assembly, maintenance, and dissociation process (de Beco et al. 2009; 2020). Treatment of this MCF-7/E-Cad-GFP cell line with the anti-estrogen Fulv for 48 h resulted in an AJ phenotype similar to the one observed in the MCF-7/vBOS cell line (Bischoff et al. 2020; Kornhuber et al. 2021) (Fig. 1A), and, in addition, influenced the E-Cad-GFP signal intensity (SI) (Fig. 1A). The SI increased in a concentration-dependent manner upon Fulv treatment (Fig. 1B, light grey) with a mean EC50 of 0.95 nM (Fig. 1C, light grey). In turn, co-treatment with increasing concentrations of 17β-Estradiol (E2) reduced the effect of Fulv on the SI again (Fig. 1B, grey) with a mean EC50 of 32.1 pM (Fig. 1C, grey).

The dependence of the SI on Fulv- or E2-mediated changes of estrogen signaling was verified by gene expression analyses of the estrogen receptor alpha (ERα) target genes *BCL2L1*, *TFF1*, *PGR*, and *AREG*. In line with results from the MCF-7/vBOS cell line (Bischoff et al. 2020), these mRNA expression levels were downregulated or upregulated in MCF-7/E-Cad-GFP cells under anti-estrogenic (Fulv) or estrogenic (Fulv + E2) conditions, whereas *ESR1* (encoding for ERα) itself was not affected (Fig. 1D). Interestingly, the expression of the *E-Cad-GFP* transgene itself clearly increased upon Fulv-treatment on the mRNA and protein level, whereas the expression level of the endogenous human E-Cad was hardly affected (Fig. 1D-F, Fig. S2A-B). The latter observation is again in line with our results from the MCF-7/vBOS cell line (Bischoff et al. 2020), whereas the interesting effect of estrogen signaling on the expression of the *E-Cad-GFP* transgene has not been described before and will be subject of future analyses. As shown by (Bischoff et al. 2020), the described changes in ERα target gene expression levels and the resulting AJ reorganization process occur at different times and differ in their kinetics. Although the authors identified several relevant cellular components that are involved in the phenotype formation process, the precise mechanisms connecting estrogen signaling activity and AJ reorganization is not yet fully understood.

Importantly, these Fulv-induced effects were indeed directly caused by inhibition of ERα activity, since specific depletion of ERα by small interfering RNAs (siRNAs) targeting *ESR1* sufficed for the formation of the AJ phenotype, the increased SI, and the elevated *E-Cad-GFP* expression levels (Fig. 1G-H; Fig. S2B). ERα activity was effectively depleted by si*ESR1* as indicated by the reduction of *AREG* and the increase of *BCL2L1* ERα target gene expression levels (Fig. S2B-C). Moreover, the application of Fulv to si*ESR1* knockdown cells only slightly further increased the SI and the *E-Cad-GFP* expression level (Fig. 1G-H). Likewise, addition of E2 could also not rescue these si*ESR1*-mediated effects (Fig. 1G-H).

Together, these data support the conclusion that the MCF-7/E-Cad-

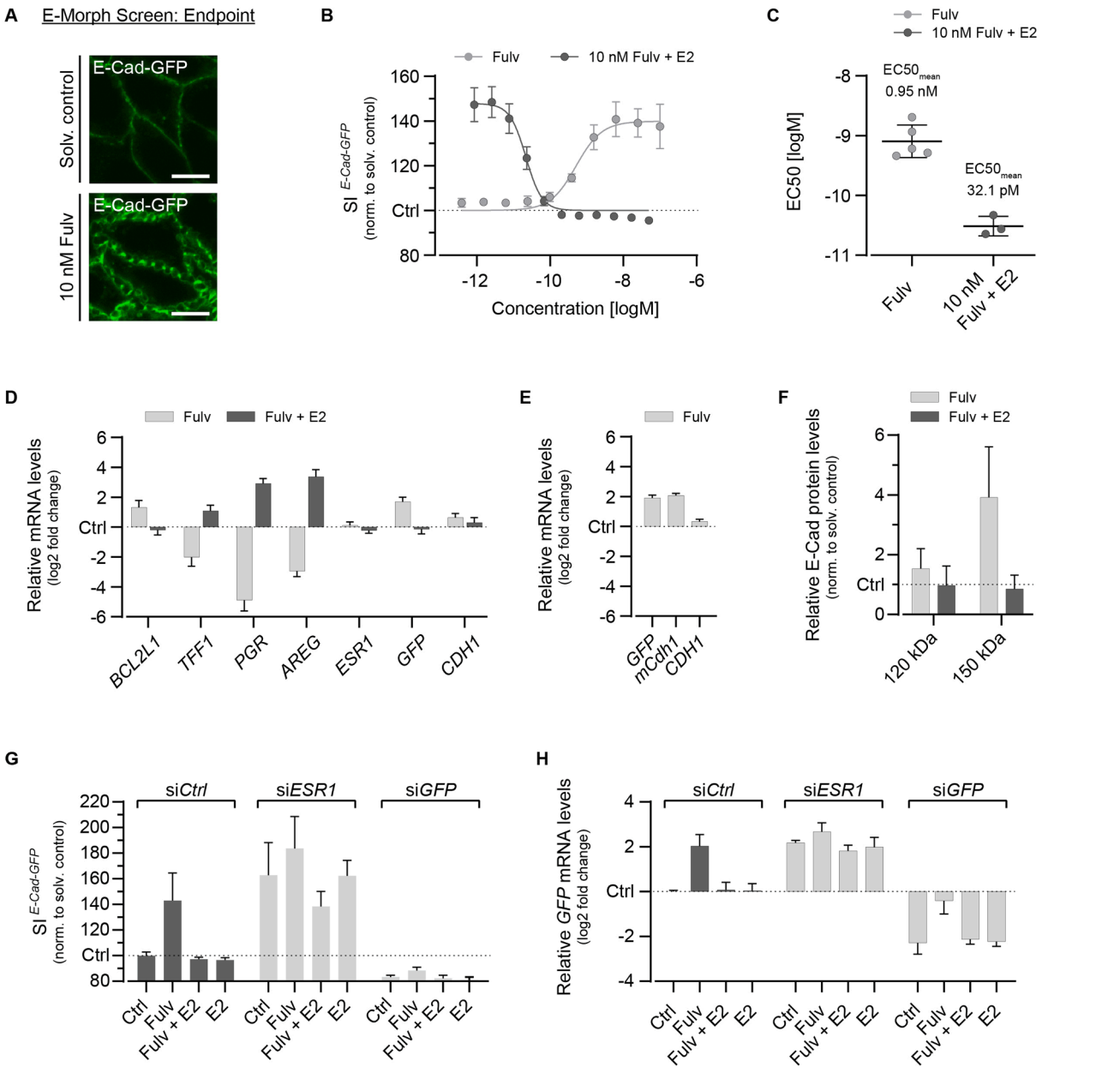

Fig. 1. Visualization and quantification of E-Cad-GFP signal intensity (SI) along the cell membrane as an endpoint for estrogenic activity in the E-Morph Screening Assay. (A) Fluorescence images showing E-Cad distribution and intercellular spacing in MCF-7/E-Cad-GFP cells upon Fulvestrant (10 nM Fulv) treatment for 48 h as compared to the solvent control. Scale bars, 10 μM. (B) Representative concentration–response curves from quantitative image analysis of E-Cad-GFP expressing cells under anti-estrogenic (Fulv titration) and estrogenic (10 nM Fulv + 17β-estradiol (E2) titration) conditions (treatment for 48 h). The plot depicts the relative E-Cad-GFP signal intensity, which increases under anti-estrogenic conditions and decreases under estrogenic conditions. Signal intensities are normalized to the solvent control (Ctrl). Non-linear fit (four parameters, variable hill slope, bottom constrained to Ctrl). Biological replicates, n = 1. Error bars, mean +/- SD from ≥ 3 technical replicate experiments. (C) Mean half-maximal concentrations ($EC50_{mean}$) derived from dose-response curves as described in Fig. 1B. Non-linear fit (four parameters, variable hill slope, bottom constrained to the solvent control). Biological replicates, n = 5 (Fulv titration, normalized to the solvent control) and n = 3 (10 nM Fulv + E2 titration, normalized to the solvent control and the 10 nM Fulv control). (D) Quantitative PCR measurement of mRNA expression levels of typical ERα target genes (*BCL2L1*, *TFF1*, *PGR*, *AREG*), *ESR1*, *GFP*, and *CDH1* under anti-estrogenic (10 nM Fulv) and estrogenic (10 nM Fulv + 10 nM E2) conditions (treatment for 48 h). Relative mRNA expression levels for each treatment condition are normalized to the solvent control (Ctrl). Biological replicates, n ≥ 3. Error bars, mean + SD. (E) Quantitative PCR measurement of mRNA expression levels of *GFP*, murine *Cdh1* and human *CDH1* under anti-estrogenic (10 nM Fulv) conditions (treatment for 48 h). Relative mRNA expression levels for each treatment condition are normalized to the solvent control (Ctrl). Biological replicates, n = 3. Error bars, mean + SD. (F) Quantification of protein expression levels of endogenous E-Cad (120 kDa) and transgenic E-Cad-GFP (150 kDa) bands from chemiluminescence western blots shown in Fig. S2A. Relative protein expression levels under anti-estrogenic (10 nM Fulv) and estrogenic (10 nM Fulv + 10 nM E2) conditions (treatment for 48 h) are normalized to the solvent control (Ctrl). Biological replicates, n = 3. Error bars, mean + SD. Loading control, Coomassie total protein staining. G) Quantification of E-Cad-GFP signal intensities from cells transfected with *ESR1* siRNA or *GFP* siRNA compared to cells transfected with scrambled control siRNA (si*Ctrl*) for 72 h. Cells were additionally grown under anti-estrogenic (10 nM Fulv) and estrogenic (10 nM Fulv + 10 nM E2 or 10 nM E2 alone) conditions (treatment for 48 h). Relative E-Cad-GFP signal intensities are normalized to cells treated with scrambled control siRNA and the solvent control (Ctrl). Biological replicates, n = 2. Error bars, mean + SD. (H) Quantitative PCR measurement of *GFP* mRNA expression levels of data shown in Fig. 1G. Relative mRNA expression levels are normalized to cells treated with scrambled control siRNA (si*Ctrl*) and the solvent control (Ctrl). Biological replicates, n = 2. Error bars, mean + SD.

GFP cell line can adequately replace the original MCF-7/vBOS cell line in the E-Morph Screening Assay and that the SI represents a novel and reliable readout for estrogenic activity. This readout further simplifies quantitative image analysis pipelines because it does not require training of a supervised machine learning algorithm for image classification such as in the original E-Morph Assay readout.

3.2. Automated high-throughput screening for estrogenic substances

Besides implementing the novel primary readout of the E-Morph Screening Assay, we adapted and optimized the cell treatment and staining procedure for automated handling of 96- and 384-well-plates using a robotic platform. To speed up the image data analysis and evaluation procedure, we further improved the automated imaging and quantitative image analysis pipeline and built an automated image data evaluation pipeline using the KNIME software (Fig. S1A-B, File S1). In this KNIME workflow, substance-related cell death or altered cell proliferation were automatically detected by counting the number of nuclei as a readout for the number of cells. In a next step, the measured SI was normalized to both the solvent control (SI = 100) and the 10 nM Fulv control (SI = 0) for cut-off-based classification of substances with potential estrogenic activity (ER activity: SI ≥ 20, i.e., representing a ≥ 20% reduction of the Fulv-mediated increase in E-Cad-GFP membrane signal intensity).

We then applied this automated HTS pipeline to screen a substance library comprising 430 toxicologically-relevant industrial chemicals, biocides, and plant protection products, as well as reference substances with already known, specific activities on different nuclear receptor

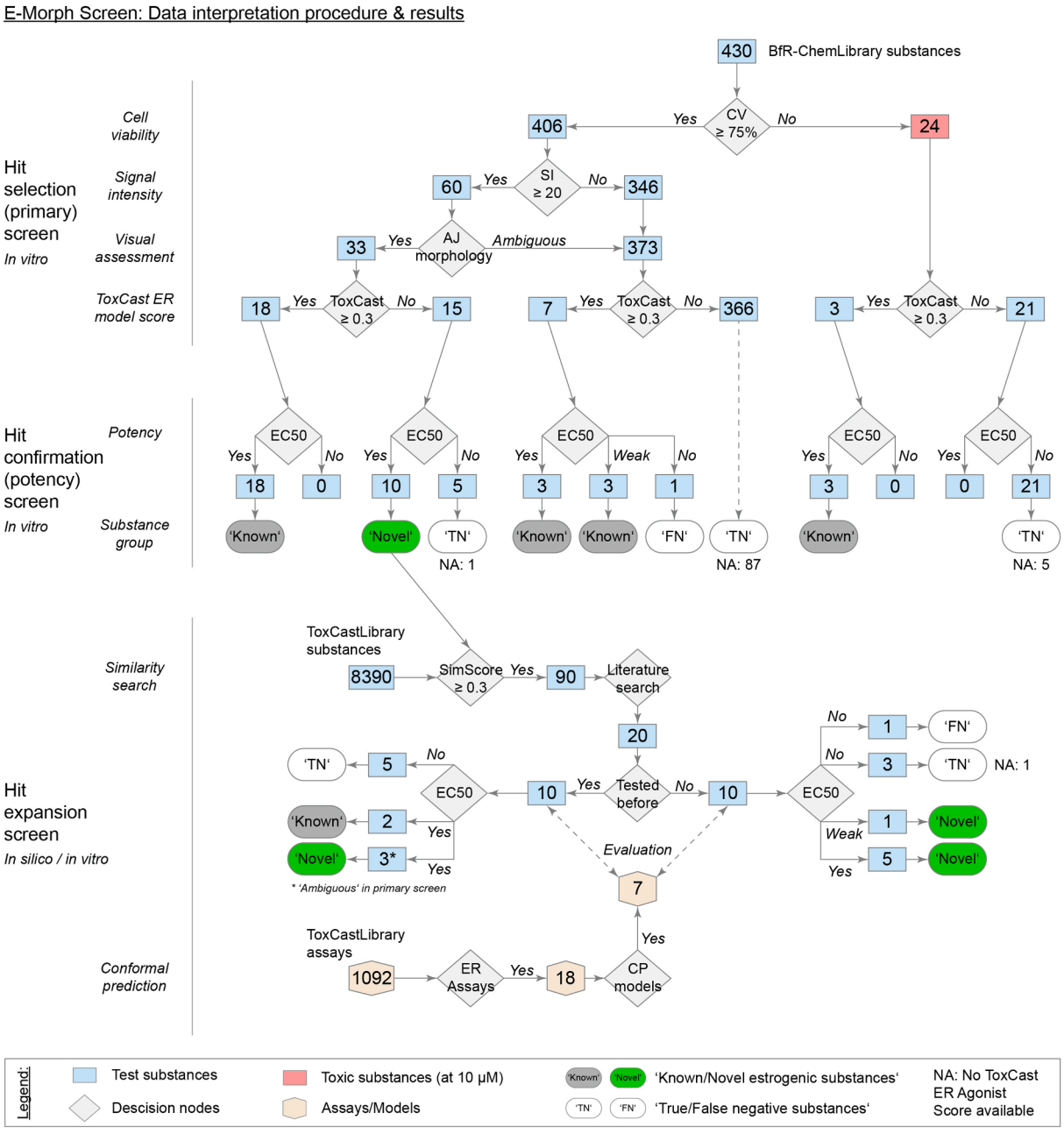

Fig. 2. E-Morph screen workflow and results. Decision tree describing the data interpretation procedure and the results of the three consecutive E-Morph screens involving *in vitro* (E-Morph Screening Assay) and *in silico* (similarity search, conformal prediction) methods to identify substances with estrogenic activity. See main text for details. Rectangular boxes, numbers of substances. Hexagonal boxes, numbers of assays/models. Oval boxes, substance groups based on comparison to the published ToxCast ER Agonist Score (see Fig. 3).

signaling pathways (Filer et al., 2014; Wetzel et al., 2017; EFSA, 2015; OECD, 2017; U.S. NIEHS, 2019). A complete list of these BfR-ChemLibrary substances and the corresponding screening results are available in Supplementary Table S1. The data interpretation procedure for identification of estrogenic substances and the screening results are summarized in Fig. 2. The quality and performance of the screens was furthermore evaluated in the KNIME workflow based on a commonly used statistical parameter, i.e., the Z'-factor (Z'). Each run achieved a Z'-factor > 0.5, which indicates a very robust HTS assay according to (Iversen et al. 2006; Zhang et al. 1999) and demonstrates the applicability of the E-Morph Screening Assay for HTS purpose. Notably, (Kornhuber et al. 2021) already compared the robustness of the selected 48 h time point with a shorter treatment period of 30 h and concluded that the effect size of the assay declined when the test chemical exposure time was reduced.

3.3. Primary screen and hit selection

In the primary screen (Fig. 2), we tested the 430 substances at a single concentration of 10 μM in the presence of 10 nM Fulv for 48 h in three independent runs. Of those, 24 substances led to a significantly reduced cell viability (CV) < 75% (Figs. 2 and 3, red) and were therefore subsequently re-tested at lower concentrations (1 pM - 30 μM) in the hit confirmation screen as described below. We identified 60 potential hit substances (SI ≥ 20), of which 33 substances clearly influenced the characteristic estrogen-dependent AJ morphology in a similar way as compared to the Estrone reactivity control (visual assessment), corresponding to an overall hit-rate of 7.7% (Fig. 2). The other 27 substances did either not clearly influence the AJ phenotype or caused rather unrelated changes in fluorescence intensity and were therefore first considered 'inactive' but flagged as 'ambiguous' in Table S1. Comparing the results for the 33 clear hit substances with the published ToxCast ER Agonist score (Browne et al. 2015; Judson et al. 2015) identified 18 substances with a ToxCast ER Agonist Score ≥ 0.3 that were considered verified actives in the primary screen (Figs. 2 and 3, grey). The ToxCast ER Agonist Score of the remaining 15 hit substances was < 0.3 or 'not available' (NA) indicating potential yet undescribed estrogenic activity (Figs. 2 and 3, green). Notably, seven of the 373 substances that were first classified as 'inactive' had a ToxCast ER Agonist Score ≥ 0.3 and were therefore included in the subsequent hit confirmation screen to be tested at higher concentrations.

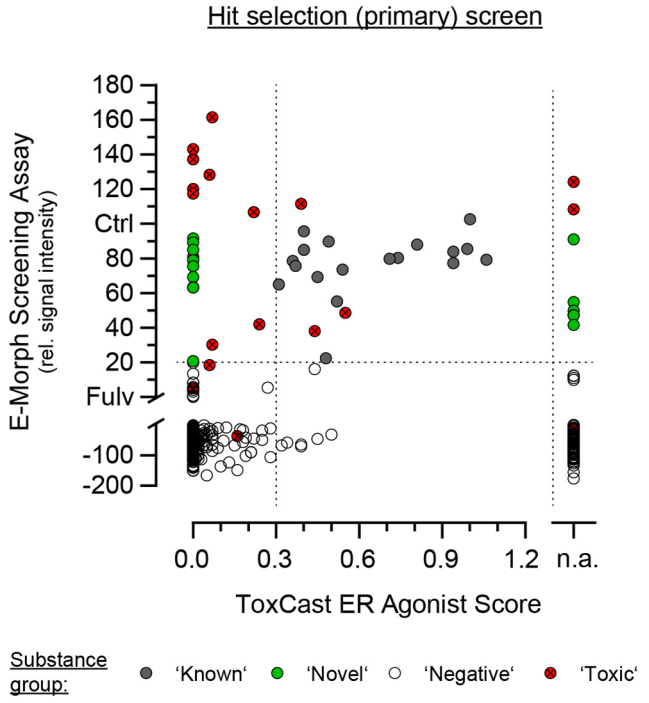

Fig. 3. Primary screening and hit selection. Relative E-Cad-GFP signal intensities of 430 test substances that were measured in the hit selection (primary) screen as compared to the published ToxCast ER Agonist Score. Cells were exposed to 10 nM Fulv + 10 μM test substance for 48 h. Each data point represents the mean relative signal intensity obtained from three independent runs. Relative E-Cad-GFP signal intensities are normalized to the solvent control (Ctrl, SI = 100) and the 10 nM Fulv control (Fulv, SI = 0). Substances that induce an increase of the relative signal intensity above the assay threshold (horizontal dashed line, SI ≥ 20) are considered as primary hit substances. Hit substances with a ToxCast ER Agonist Score ≥ 0.3 (vertical dashed line) are depicted in grey color and assigned to the group of 'Known estrogenic substances'. Hit substances with a ToxCast ER Agonist Score < 0.3 or 'not available' (NA) (vertical dashed line) are highlighted in green color and assigned to the group of potential 'Novel estrogenic substances'. Test substances leading to a cell viability (CV) < 75% are indicated in red color and assigned to the group of 'Toxic substances'. Data from substances that were excluded after visual assessment of images are not displayed. Biological replicates, n = 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Hit confirmation screen and potency determination

In the hit confirmation screen (Fig. 2), we re-tested the 33 primary hit substances (SI ≥ 20), the 24 substances displaying cytotoxicity at 10 μM, and the seven 'inactive' substances with a ToxCast ER agonist score ≥ 0.3 (in total 64 substances) at multiple concentrations ranging from 1 pM to 30 μM in the presence of 10 nM Fulv for 48 h in multiple independent runs. A clear concentration-dependent estrogenic activity was detected for 28 out of the 33 primary hit substances (SI ≥ 20), three out of the 24 'cytotoxic' substances (CV < 75%) when tested at concentrations < 10 μM, and six out of the seven 'inactive' substances (ToxCast ER agonist score ≥ 0.3) when tested at concentrations > 10 μM. The remaining substances did not show a clear activity in the tested concentration range. These data show that the results of the primary screen and the hit confirmation screen were concordant to a large extent and highlight the need of testing substances in a wide concentration range to increase the hit rate. Based on the respective ToxCast ER Agonist Score, we grouped the active substances as 'Known estrogenic substances' (≥0.3; 27 substances) (Fig. 2, grey; Table 1) or potential 'Novel estrogenic substances' (<0.3 or not available ('NA'); 10 substances) (Fig. 2, green; Table 2). Accordingly, the inactive substances were grouped as 'True negative substances' ('TN') (<0.3; 299 substances), 'False negative substances' ('FN') (≥0.3; 1 substance), or 'NA' if no ToxCast ER Agonist Score was available (93 substances) (Fig. 2, white; Table S1).

Using the concentration–response data that was collected in the hit confirmation screen, we further determined the potencies (EC50) and relative ER bioactivities (logEC50 normalized to 17-alpha-Ethinylestradiol) for 24 out of the 27 'Known estrogenic substances' (Fig. 2, {'Known','Yes'}; Fig. 4A; Table 1), which correlated well with the respective ToxCast ER Agonist Scores (r = +0.95) (Fig. 4A-B; Table 1). No EC50 values could be determined for the remaining three substances, including Tamoxifen, which showed only weak estrogenic activity in the hit confirmation screen (Fig. 2, {'Known','Weak'}; Table 1). The weak estrogenic activity of Tamoxifen might reflect its partial agonistic function (Jordan 1977), which is further supported by its estrogenic activity in the uterotrophic bioassay in rodents (Kleinstreuer et al. 2016). Overall, the measured activities were also concordant with the CERAPP ER consensus model predictions (Table 1). Among the substances with a ToxCast ER Agonist Score ≥ 0.3, only 2,4′-DDT may have been a potential false negative substance in the E-Morph Screen (Fig. 2, 'FN'; Table 1). This particular chemical appears to be difficult to detect in the tested concentration range as it also showed weak or no activity in other ER testing systems according to the Integrated Chemical Environment database (Bell et al. 2020; Bell et al. 2017) of the U.S. National Toxicology Program.

Table 1
Screening results for the group of 'Known estrogenic substances' compared to published *in silico* ER model data from the U.S. EPA.

Chemical name	CAS No.	U.S. EPA *in silico* ER models		E-Morph Screening Assay	Hit confirmation screen				
		ToxCast ER Agonist Score[a)]	CERAPP ER Agonist Model[b)]	Substance group	Potency [M]			ER Bioactivity [rel. LogEC50]	Comment
					EC50	SD	n		
Diethylstilbestrol	56-53-1	0.94	active	Known	7.97E-10	6.99E-10	3	1.03	active
Beta-Estradiol	50-28-2	0.94	active	Known	9.00E-10	1.30E-10	3	1.03	active
Hexestrol	84-16-2	0.99	active	Known	1.05E-09	2.59E-10	4	1.02	active
Ethinyl Estradiol	57-63-6	1.00	active	Known	1.56E-09	3.04E-10	3	1.00	active
Mestranol	72-33-3	0.74	active	Known	2.26E-09	7.13E-10	4	0.98	active
Estrone	53-16-7	0.81	active	Known	3.18E-09	2.83E-09	4	0.96	active
Alpha-Estradiol	57-91-0	1.06	active	Known	5.50E-09	1.74E-09	4	0.94	active
Zearalenone	17924-92-4	0.71	active	Known	1.66E-07	6.01E-08	4	0.77	active
Bisphenol AF	1478-61-1	0.55	active	Known	2.96E-07	1.47E-07	3	0.74	active
Trenbolone-Dea	10161-33-8	0.48	active	Known	4.30E-07	1.70E-07	3	0.72	active
Genistein	446-72-0	0.54	active	Known	4.35E-07	4.00E-07	4	0.72	active
Norethisterone	68-22-4	0.52	active	Known	7.71E-07	4.22E-07	4	0.69	active
Bisphenol B	77-40-7	0.49	active	Known	1.07E-06	5.16E-07	4	0.68	active
5alpha-Androstan-17beta-OL-3ON	521-18-6	0.40	active	Known	1.28E-06	1.59E-06	4	0.67	active
Nonylphenol techn Gemisch	84852-15-3	0.44	active	Known	1.39E-06	1.77E-07	2	0.67	active
Dehydroisoandrosterone	53-43-0	0.37	active	Known	1.54E-06	6.70E-07	4	0.66	active
4-tert.-Octylphenol	140-66-9	0.39	active	Known	1.80E-06	5.59E-07	2	0.65	active
Bisphenol A	80-05-7	0.45	active	Known	2.23E-06	4.10E-07	3	0.64	active
Biochanin A	491-80-5	0.36	active	Known	3.09E-06	1.05E-06	4	0.63	active
Daidzein	486-66-8	0.44	active	Known	3.10E-06	9.26E-07	2	0.63	active
2,2',4,4'-Tetrahydroxybenzophenon	131-55-5	0.40	active	Known	4.26E-06	2.32E-06	4	0.61	active
17alpha-Hydroxyprogesterone	68-96-2	0.34	active	Known	8.22E-06	4.64E-06	2	0.58	active
Levonorgestrel	797-63-7	0.39	active	Known	8.79E-06	2.99E-06	2	0.57	active
Apigenin	520-36-5	0.31	active	Known	2.16E-05	2.16E-05	4	0.53	active
1,1,1-Tris(4-hydroxyphenyl) ethane	27955-94-8	0.32	active	Known	NA	NA	1	NA	weakly active at 30 µM (22,3 % rescue)
2,4'-DDT Lösung	789-02-6	0.39	active	FN	NA	NA	1	NA	negative
17a-Methyltestosterone	58-18-4	0.50	active	Known	NA	NA	1	NA	weakly active at 10 µM (21,3 % rescue)
Tamoxifen	10540-29-1	0.45	inactive	Known	NA	NA	1	NA	weakly active at 10 µM (26,6 % rescue)

Overall classifications and potencies of 28 substances with a ToxCast ER Agonist Score $\geq$ 0.3. EC50, mean potency from multiple independent runs (n). SD, standard deviation. ER Bioactivity, potency (logEC50) normalized to 17-alpha-Ethinylestradiol (1.00). NA, not available/not applicable. FN, false negative substance.
[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015)
[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016)

Potencies (EC50) and relative ER bioactivities were also determined for all 10 substances that were assigned to the group of 'Novel estrogenic substances' (Fig. 2, 'Novel'; Fig. 4C; Table 2). The calcium chelating agent EDTA was considered a false positive substance because of its known interference with the Ca + -dependent E-Cad activity and therefore excluded from further analysis. The two most potent substances in this hit class were the pesticide Hexythiazox (insecticide) with an EC50 of 10 nM, for which no estrogenic activity has been described before, and the progestin Norethisterone acetate (NETA) with an EC50 of 625 nM. The remaining substances showed weaker activities in the µM range. The estrogenic activities of NETA, Nandrolone (androgen and anabolic steroid (AAS)), Phloretin (flavonoid), and Bisphenol F (industrial chemical), for which no ToxCast ER Agonist Score was available, were consistent with previous studies (Branham et al. 2002; Chwalisz

Table 2
Screening results for the group of 'Novel estrogenic substances' compared to published *in silico* ER model data from the U.S. EPA.

Chemical name	CAS No.	U.S. EPA *in silico* ER models		E-Morph Screening Assay	Hit confirmation screen				
		ToxCast ER Agonist Score [a)]	CERAPP ER Agonist Model [b)]	Substance group	Potency [M]			ER Bioactivity [rel. LogEC50]	Comment
					EC50	SD	n		
Hexythiazox	78587–05-0	0.00	inactive	Novel	1.01E-08	3.94E-09	4	0.91	active
Norethindrone acetate (NETA)	51–98-9	NA	active	Novel	6.25E-07	4.62E-07	6	0.70	active
EDTA iron(III) sodium salt	15708–41-5	0.00	inactive	Novel	1.10E-06	7.34E-07	3	0.68	active
Nandrolone	434–22-0	NA	active	Novel	2.04E-06	1.38E-06	6	0.65	active
Phloretin	60–82-2	NA	active	Novel	3.36E-06	1.41E-06	5	0.62	active
2,4,6-Tri-*tert*-butylphenol (TTBP)	732–26-3	0.00	inactive	Novel	4.73E-06	3.12E-06	4	0.60	active
Bisphenol F	620–92-8	NA	active	Novel	4.79E-06	1.41E-06	5	0.60	active
Diuron	330–54-1	0.00	inactive	Novel	6.04E-06	2.91E-06	3	0.59	active
Azoxystrobin	131860–33-8	0.00	inactive	Novel	6.34E-06	3.68E-06	6	0.59	active
Zineb	12122–67-7	NA	inactive	Novel	8.71E-05	8.48E-05	2	0.46	active

Overall classifications and potencies of 10 substances that were active in the E-Morph Screening Assay with a ToxCast ER Agonist Score = 0.00 or 'not available' (NA). EC50, mean potency from multiple independent runs (n). SD, standard deviation. ER Bioactivity, potency (logEC50) normalized to 17-alpha-Ethinylestradiol (1.00). NA, not available/not applicable.
[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015)
[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016)

et al. 2012; Rochester and Bolden 2015; Sirianni et al. 2012), results from ER transactivation screening assays (Table S4), as well as the CERAPP ER consensus model (Table 2). Nandrolone (19-nortestosterone) was further shown to be active in the STTA and VM7Luc ER transactivation assays of OECD TG 455 (OECD 2016). For 2,4,6-TTBP (industrial chemical), Azoxystrobin (fungicide), Hexythiazox, and Diuron (herbicide), the ToxCast ER Agonist Score was 0.00 and they were also inactive in the relevant ER screening assays (Table S4) as well as the CERAPP ER consensus model (Table 2). Furthermore, these substances were also classified as 'non-binder' in the FW and CERI ER binding assays of OECD TG 493 (OECD 2015). In addition, for the fungicide Zineb (Zink-ethylen-1,2-bis-dithiocarbamat) no conclusive data demonstrating estrogenic activity was available, yet (Table S4).

These partially discordant results between the E-Morph Screening Assay, the ToxCast ER HTS data, and the *in silico* ER models (ToxCast and CERAPP) can have various reasons (including false positive results), but may also reflect that, in contrast to the simplifying ER HTS assays conducted in the ToxCast project, the functional E-Morph Screening Assay integrates multiple interacting cellular pathways. On the one hand, it therefore provides a more complete picture of relevant cellular mechanisms mediating estrogen-dependent effects. On the other hand, integration of multiple mechanistic events or cellular signaling pathways in a single assay increases the degrees of freedom for possible modes of action of test substances and necessitates running secondary assays to confirm substance-specific effects on distinct signaling pathways.

3.5. Verification of 'Novel estrogenic substances'

In order to verify the nine (excluding EDTA) potential 'Novel estrogenic substances' (Table 2), we first determined the mRNA expression levels of the ERα target genes *BCL2L1*, *TFF1*, *PGR*, and *AREG* along with *ESR1* and *CDH1*. Cells were exposed to each test substance at a concentration of 10 μM in the presence of 10 nM Fulv for 48 h and the effects were compared to the mRNA expression profiles under anti-estrogenic (Fulv) and estrogenic (Fulv + E2) conditions (Fig. 5A; Fig. S2D). Hexythiazox and NETA showed the most similar expression profiles when compared to the E2 reference substance, which was in line with the high potency that was measured in the E-Morph Screening Assay. Nandrolone, Phloretin, and Diuron also showed an estrogenic expression profile, albeit to a weaker extent. Bisphenol F slightly inhibited the Fulv effect, particularly for *TFF1* expression. Importantly, these effects on gene expression profiles could be confirmed when cells were only exposed to these test substances without Fulv co-treatment (Fig. 5B; Fig. S2E). Interestingly, Bisphenol F showed the strongest effect in this case. The expression profiles of the phenol 2,4,6-TTBP and the pesticide Zineb did not support an estrogenic effect neither in the competitive treatment nor in the single treatment scenario (Fig. 5A-B; Fig. S2D-E). The gene expression pattern of Azoxystrobin rather showed some surprising anti-estrogenic effect, particularly for *TFF1* and *PGR*, when applied to cells without Fulv (Fig. 5B; Fig. S2E).

To characterize the underlying mechanism of action of the potential 'Novel estrogenic substances' (Table 2), we next performed a 'pBIND-ERα vector assay', which allows the identification of substances that directly bind to the ERα ligand binding domain. In this assay, MCF-7/E-Cad-GFP cells were transiently transfected with both the reporter and control vector plasmids and subsequently treated with each test substance at 10 μM for 48 h (Fig. 5C). The results were very similar to the detected gene expression profiles. Hexythiazox and NETA showed the highest activity in this assay, whereas Nandrolone, Phloretin, Diuron, and Bisphenol F showed weaker effects. Again, an estrogenic activity at the level of ERα binding could not be identified for 2,4,6-TTBP, Zineb, and Azoxystrobin.

Together, these secondary assay data support the detected estrogenic activity of Hexythiazox, NETA, Nandrolone, Phloretin, Diuron, and Bisphenol F from the group of potential 'Novel estrogenic substances' that were identified by the E-Morph Screening Assay (Fig. 2; Table 2). These data further underline the importance of running secondary assays to identify potential false positive substances, i.e., 2,4,6-TTBP, Zineb, and Azoxystrobin. These substances might act on E-cadherin or AJs in an estrogen-independent manner that will be addressed in future analyses.

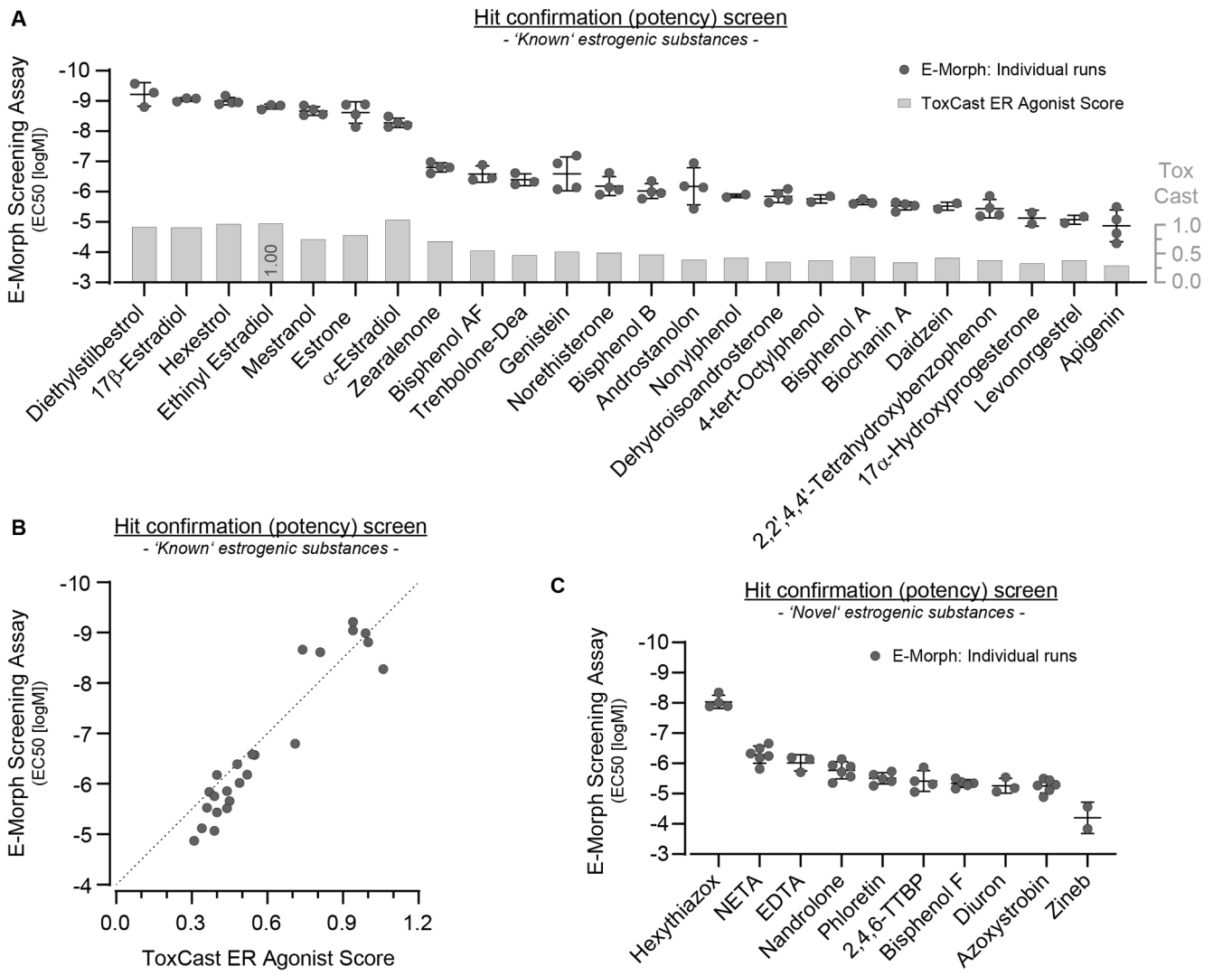

Fig. 4. Hit confirmation screening and potency determination. A) Potencies (half-maximal concentrations, EC50) of 24 'Known' hit substances that were active in the hit confirmation (potency) screen with a ToxCast ER Agonist Score ≥ 0.3. Each data point represents an individual run. The three 'Known' hit substances 1,1,1-Tris-(4-hydroxyphenyl)-ethan, Methyltestosterone, and Tamoxifen showed only weak activity (no EC50 value could be determined) and are therefore not shown. Biological replicates, n ≥ 3. Error bars, mean +/- SD. (B) Correlation between the potencies (half-maximal concentrations, EC50) of 24 'Known' hit substances obtained from the E-Morph Assay and the published ToxCast ER Agonist Score. Each data point represents the mean of the relative bioactivities obtained from individual runs shown in Fig. 4A. The contour line indicates full correlation. Pearson r = +0.95. (C) Potencies (half-maximal concentrations, EC50) of 10 'Novel' hit substances that were active in the hit confirmation (potency) screen with a ToxCast ER Agonist Score = 0.00 or 'not available' (NA). Each data point represents an individual run. Biological replicates, n ≥ 3. Error bars, mean +/- SD.

3.6. Identification of structurally similar substances and hit expansion screening

Based on the assumption that structurally similar substances can interact with similar targets (Bender and Glen 2004), we performed an *in silico* similarity search against the substances for which ToxCast assay data was available (Fig. 2). The chemical structures of the nine (excluding EDTA) potential 'Novel estrogenic substances' (Table 2) were used as input for the identification of other structurally similar substances (Table S2; File S2). Based on the resulting similarity scores (Tanimoto index) and literature search, we selected a final set of 20 similar substances and measured their potential estrogenic activity and potency in the E-Morph Screening Assay (Fig. 6; Table 3). Notably, we set a relatively low global Tanimoto cut-off (>0.3) for inclusion of similar substances into the hit expansion screening in order to account for the diverse structural complexities of the input substances. Furthermore, the number of selected similar substances per input substance varied because of the composition of the ToxCast substance library. For example, the ToxCast database contains assay data for many different bisphenols with relatively high similarity scores (>0.5) to Bisphenol F, but no data for substances that are similar to the pesticide Hexythiazox with a score above 0.4 (Table 3; Table S2; File S2).

Of the 20 similar substances, 10 substances had already been tested in the primary or hit confirmation screens (Fig. 2; Table 3) and their re-testing in the hit expansion screening (5 active, 5 inactive) provided concordant results (Table S1). Interestingly, three of these substances (4,4′-Dihydroxybiphenyl (92–88-6), 4,4′-Dihydroxybenzophenone (611–99-4), Triclocarban (101–20-2)) were initially considered as 'ambiguous' by visual inspection in the primary screen at 10 μM (Fig. 2; Table S1). However, based on the hit expansion screening of a wider concentration range, these substances could now be re-assignment to the group of 'Novel estrogenic substances' (Table 3; Table S1). Hence, a visual inspection of the AJ phenotype at a single concentration also bears the risk of misinterpretation of substance effects, particularly at concentrations near the cytotoxic range. Of the remaining 10 newly tested substances, six substances were active and four substances were

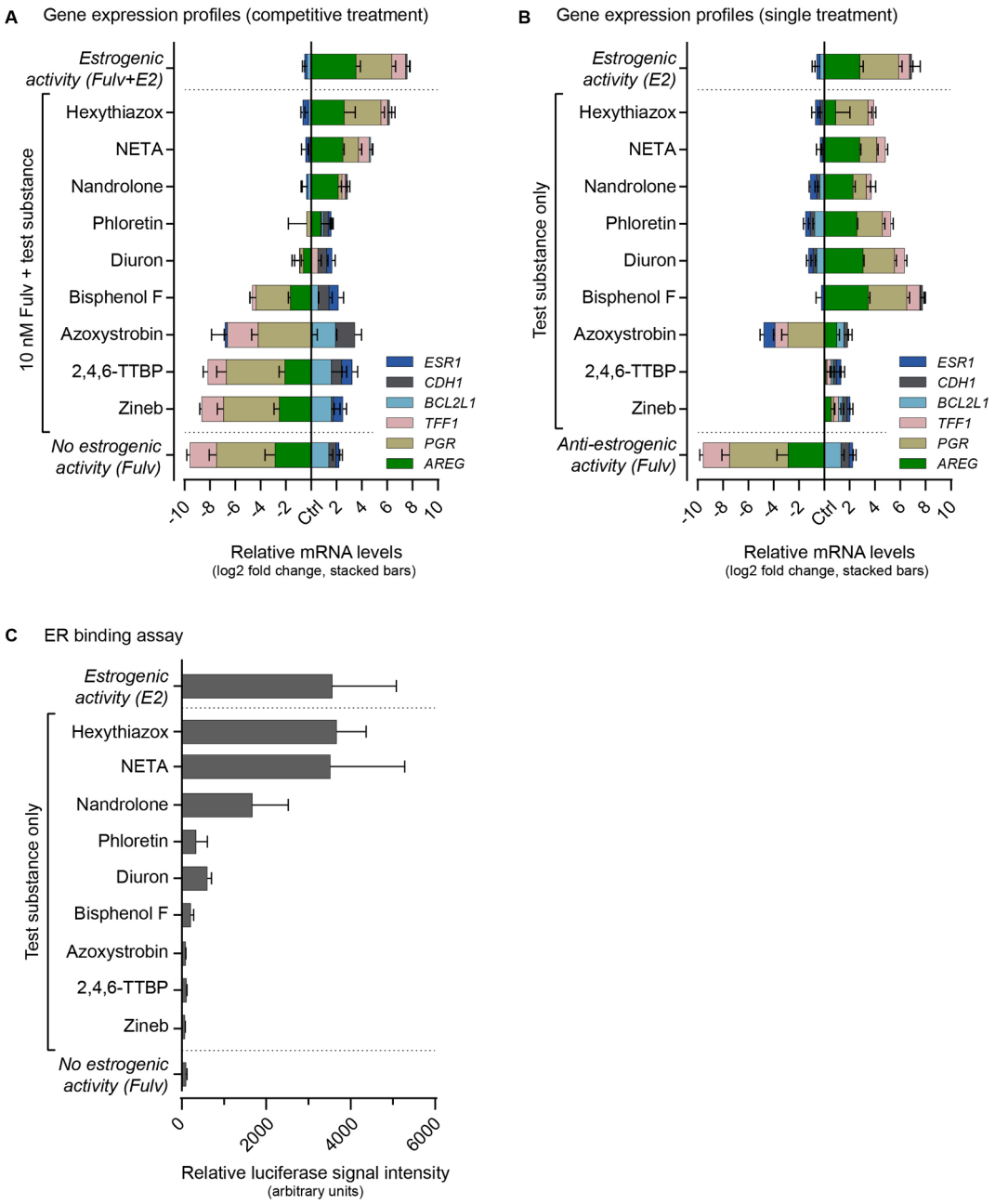

Fig. 5. Verification of 'Novel' hit substances. (A) Gene expression profiles from quantitative PCR measurements of *ESR1*, *CDH1*, and typical ERα target genes (*BCL2L1*, *TFF1*, *PGR*, *AREG*). Cells were exposed to 10 nM Fulv + 10 μM test substance (competitive treatment) for 48 h. Measurements from estrogenic (10 nM Fulv + 10 nM E2, top) and anti-estrogenic (10 nM Fulv, bottom) conditions provide reference gene expression profiles. Relative mRNA expression levels are normalized to cells treated with the solvent control (Ctrl). Biological replicates, n = 3. (B) Gene expression profiles from quantitative PCR measurements of *ESR1*, *CDH1*, and typical ERα target genes (*BCL2L1*, *TFF1*, *PGR*, *AREG*). Cells were exposed to 10 μM test substance only (single treatment) for 48 h. Measurements from estrogenic (10 nM E2, top) and anti-estrogenic (10 nM Fulv, bottom) conditions provide reference gene expression profiles. Relative mRNA expression levels are normalized to cells treated with the solvent control (Ctrl). Biological replicates, n = 3. C) Relative luciferase signal intensities obtained from an ERα reporter gene assay. Cells were co-transfected for 72 h with a pBIND-ERα expression vector (Gal4-DBD fused to ERα-LBD, Renilla luciferase) and a target vector expressing a UAS-controlled Firefly luciferase. Cells were exposed to 10 μM test substance only for 48 h. The detected Firefly luminescence was normalized to Renilla luminescence. Relative signal intensities from estrogenic (10 nM E2, top) and anti-estrogenic (10 nM Fulv, bottom) conditions provide reference measurements. Biological replicates, n = 3. Error bars, mean + SD.

inactive (Fig. 2; Table 3, Table S1). Interestingly, Norgestrel, a racemate of D-Norgestrel and L-Norgestrel/Levonorgestrel enantiomers (Kuhl 2005), was inactive in the hit expansion screen despite a ToxCast ER Agonist Score of 0.39 (Table 3, 'FN'), whereas the known active enantiomer Levonorgestrel by itself was active in our assay (Table 3, 'Known'). Thus, selecting candidate substances based on structural similarities to our primary hits, resulted in the identification of in total nine additional 'Novel estrogenic substances'. The measured estrogenic activities for most of these substances were supported by the CERAPP ER consensus model (Table 3).

Together, these screening data provide strong support for an estrogenic activity of NETA, Nandrolone, Phloretin, and Bisphenol F and demonstrate that the combination of *in silico* similarity search and *in vitro* testing supports the identification of estrogenic activities.

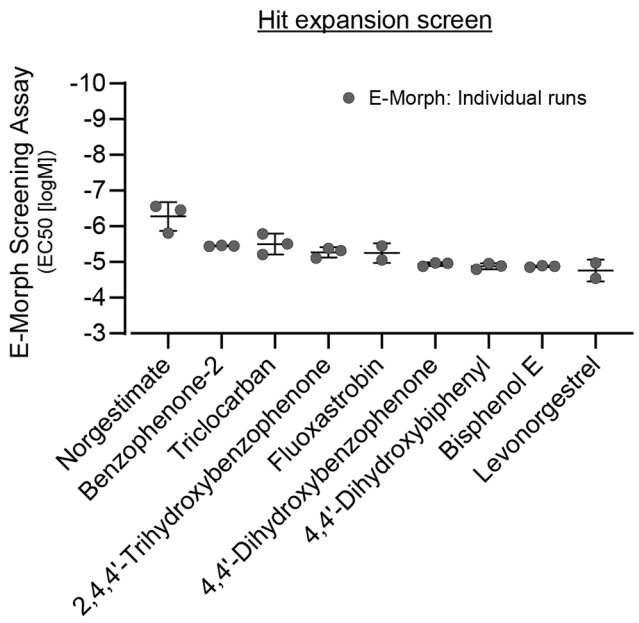

Fig. 6. Hit expansion screening and potency determination. Potencies (half-maximal concentrations, EC50) of nine additional hit substances that were active in the hit expansion screen. Each data point represents an individual run. Biological replicates, n = 3. Error bars, mean +/- SD.

Although, potential false positive results, like 2,4,6-TTBP, Zineb, or Azoxystrobin, are inevitable, the identification of Hexythiazox and Diuron as potential estrogenic substances seems to be relevant and warrant further investigation.

3.7. Docking of bisphenols into ERα

Even though the primary hit substance Bisphenol F shares a very similar chemical structure with 4-Benzylphenol (Fig. 7, File S2), the latter was inactive in the E-Morph Screening Assay (Table 3; Table S1). This was particularly surprising since this substance had the highest Tanimoto similarity score of 0.78 as compared to three additional substances with similarity scores between 0.49 and 0.56 that were active in the E-Morph Assay. Here, it needs to be considered that terminal functional group differences in molecules often have less influence on the calculated similarity than central atom changes (because of the nature of the calculation of the circular environments, such as in the Morgan fingerprint). The different activities could neither be explained using diverse other molecular fingerprints (MorganCount, MACCS, pharmacophore fingerprints), which were very similar for these bisphenols (Table S3). In scientific literature, such phenomena are often described as activity cliffs (Maggiora et al. 2014) in which, e.g., changes of one atom or functional group in otherwise very similar molecules can lead to a notable difference in activity. To better understand this effect, we performed a docking analysis (Brooijmans and Kuntz 2003) and investigated potential variations in interaction patterns of the bisphenols with ERα using the @TOME-2 web-server according to (Delfosse et al. 2012). The docking results showed an apparent difference in the binding of 4-Benzylphenol to ERα (Fig. 7, red arrow) as compared to the other four tested bisphenols. The latter four substances are all capable of forming undirected interactions with the hydrophobic pocket core, while being anchored on both sides through hydrogen bonds (see pharmacophoric interactions in Fig. 7). Hence, the inactivity of 4-Benzylphenol in the E-Morph Screening Assay could be explained by the absence of an important hydrogen bond to histidine H524 (because of a missing hydroxy (–OH) group), which may significantly reduce the binding capacity of 4-Benzylphenol to ERα.

3.8. Generation of in silico models for prediction of estrogenic activities in HTS approaches

Ideally, these kinds of structural analyses will eventually lead to the development of highly predictive (Q)SAR tools to identify estrogenic activities in environmental chemicals based on structural fingerprints. As a proof-of-concept for the application of *in silico* prediction models in HTS approaches, we used ToxCast ER screening data to build seven *in silico* prediction models based on the conformal prediction (CP) framework (Morger et al. 2020; Norinder et al. 2014; Vovk et al. 2005) for the relevant mechanistic events of estrogen signaling, i.e., ER binding, ER dimerization, regulation of gene expression, and cell proliferation (Table 4), that are also included in the ToxCast ER Agonist Model (Browne et al. 2015). We trained these CP ER models on all binary readouts (active/inactive) of the corresponding *in chemico* and *in vitro* ER screening assays that were conducted in the ToxCast project. A fivefold cross-validation was employed to assess the performance of each model (Table 4). The benefit of the CP method over simpler similarity search is thereby two-fold. First, machine learning (ML) models are statistical models that relate a set of structural descriptors of a chemical compound to its biological activity. Thus, the CP model learns which features in the molecule contribute more (or less) to the outcome, whereas a simpler similarity search treats all features the same. In other words, while similarity search looks for more obvious similarities between molecules, CP may detect more hidden, and also non-linear, relationships. Second, CP is built on top of a ML framework and adds a calibration step. This allows for monitoring the reliability of the predictions more closely, thus providing a measure of confidence in the prediction per molecule.

All CP ER models were valid at the 0.2 significance level (validity $\geq$ 0.8), i.e., making<20% prediction errors when considering 'single class' (active or inactive) and 'both class' (active and inactive) predictions (Table 4). This high mean validity of 0.85 $\pm$ 0.01 indicates that the models were well calibrated and can therefore be reliably applied to new data. The mean efficiency, i.e., the fraction of single class predictions made by the models, was 0.39 $\pm$ 0.12 (Table 4) and notably lower than compared to other CP ER models described before (Ji et al. 2018; Norinder et al. 2016). This can be a consequence of the use of the additional normalizer regression model and prior equal size sampling of the proper training and calibration set, which was shown to improve the prediction performance on external data (Morger et al. 2020). The CP ER models had a mean accuracy, i.e., the fraction of correct single class predictions made by the models, of 0.71 $\pm$ 0.10 (Table 4). Regarding the class-wise evaluation, the mean accuracy for prediction of the active class was rather high with an average of 0.83 $\pm$ 0.03, whereas the mean accuracy for prediction of the inactive class was slightly lower with an average of 0.67 $\pm$ 0.13 (Table 4). The reduced mean accuracy for inactive class predictions was mainly caused by two rather weakly performing endpoint models (aeid_788 and aeid_2) (Table 4, italics). Nevertheless, the overall results show that the individual CP ER models can reliably predict agonistic ER activity, especially since the focus of this study is to detect active substances.

We then applied the seven CP ER models to classify the nine (excluding EDTA) potential 'Novel estrogenic substances' (Table 2) from primary and hit confirmation screening as well as the selected 20 structurally similar substances from the hit expansion screening (Table 3). The respective p-values, which describe the certainty of the active/inactive predictions of the individual models, are summarized in Table S4. To facilitate direct comparison of the CP ER model classifications with the E-Morph Screening Assay results, we converted the seven individual CP ER model predictions into an overall 'consensus prediction' by applying a 'majority rule' principle (Table 5; Table S4). The *in chemico* and *in vitro* test results of the seven corresponding ER screening assays included in the ToxCast project were converted in the same way to obtain an overall 'consensus test result' for each individual substance (Table 5; Table S4). Notably, for some of the 29 test

Table 3
Screening results for the hit expansion substances compared to published *in silico* ER model data from the U.S. EPA.

Chemical name	CAS No.	US EPA *in silico* ER models		Similarity search	E-Morph Screening Assay	Hit expansion screen					
		ToxCast ER Agonist Score [a]	CERAPP ER Agonist Model [b]	Similarity score	Substance group	Potency [M] EC50	SD	n	ER Bioactivity [rel. LogEC50]	Comment	Tested in primary screen
Hexythiazox	***78587–05-0***	***0.00***	***inactive***	***1.00***	***Novel***	***1.01E-08***	***3.94E-09***	***4***	***0.91***	***active***	***Y***
Iprodion	36734–19-7	0.00	inactive	0.36	TN	NA	NA		NA	inactive	Y
Norethindrone acetate (NETA)	***51–98-9***	***NA***	***active***	***1.00***	***Novel***	***6.25E-07***	***4.62E-07***	***6***	***0.70***	***active***	***Y***
Ethynodiol diacetate	297–76-7	NA	active	0.73	Novel	< 1.37E-08	< 1.37E-08	3	NA	active	–
Norgestrel	6533–00-2	0.39	active	0.55	FN	NA	NA		NA	inactive	–
Levonorgestrel	797–63-7	0.39	active	0.55	Known	1.97E-05	1.27E-05	2	0.53	active	Y
Norgestimate	35189–28-7	NA	active	0.51	Novel	7.27E-07	7.14E-07	3	0.70	active	–
Nandrolone	***434–22-0***	***NA***	***active***	***1.00***	***Novel***	***2.04E-06***	***1.38E-06***	***6***	***0.65***	***active***	***Y***
Norgestrel	6533–00-2	0.39	active	0.60	FN	NA	NA		NA	inactive	–
Levonorgestrel	797–63-7	0.39	active	0.60	Known	1.97E-05	1.27E-05	2	0.53	active	Y
Phloretin	***60–82-2***	***NA***	***active***	***1.00***	***Novel***	***3.36E-06***	***1.41E-06***	***5***	***0.62***	***active***	***Y***
Benzophenone-2	131–55-5	0.40	active	0.43	Known	3.55E-06	1.00E-07	3	0.62	active	Y
2,4,4′-Trihydroxybenzophenone	1470–79-7	NA	active	0.42	Novel	5.60E-06	1.98E-06	3	0.60	active	–
Diphenolic acid	126–00-1	0.17	active	0.38	Novel	NA	NA		NA	active at ≥ 30 μM	–
2,4,6-Tri-tert-butylphenol (TTBP)	***732–26-3***	***0.00***	***inactive***	***1.00***	***Novel***	***4.73E-06***	***3.12E-06***	***4***	***0.60***	***active***	***Y***
Butylhydroxytoluene	128–37-0	0.00	inactive	0.72	TN	NA	NA		NA	inactive	Y
2,5-Di-*tert*-butylhydroquinone	88–58-4	0.00	inactive	0.60	TN	NA	NA		NA	inactive	–
Bisphenol F	***620–92-8***	***NA***	***active***	***1.00***	***Novel***	***4.79E-06***	***1.41E-06***	***5***	***0.60***	***active***	***Y***
4-Benzylphenol	101–53-1	NA	active	0.78	NA	NA	NA		NA	inactive	Y
4,4′-Dihydroxybiphenyl	92–88-6	NA	active	0.56	Novel	1.34E-05	2.59E-06	3	0.55	active	Y
4,4′-Dihydroxybenzophenone	611–99-4	NA	active	0.49	Novel	1.17E-05	1.42E-06	3	0.56	active	Y
Bisphenol E	2081–08-5	NA	active	0.49	Novel	1.34E-05	6.56E-07	3	0.55	active	–
Diuron	***330–54-1***	***0.00***	***inactive***	***1.00***	***Novel***	***6.04E-06***	***2.91E-06***	***3***	***0.59***	***active***	***Y***
Linuron	330–55-2	0.00	inactive	0.67	TN	NA	NA		NA	inactive	Y
Swep	1918–18-9	NA	inactive	0.59	NA	NA	NA		NA	inactive	–
Troclocarban	101–20-2	0.00	inactive	0.52	Novel	3.65E-06	2.33E-06	3	0.62	active	Y
Azoxystrobin	***131860–33-8***	***0.00***	***inactive***	***1.00***	***Novel***	***6.34E-06***	***3.68E-06***	***6***	***0.59***	***active***	***Y***
Picoxystrobin	117428–22-5	0.00	inactive	0.45	TN	NA	NA		NA	inactive	–
Fluoxastrobin	361377–29-9	0.00	inactive	0.33	Novel	6.23E-06	3.72E-06	2	0.59	active	–
Zineb	***12122–67-7***	***NA***	***inactive***	***1.00***	***Novel***	***8.71E-05***	***8.48E-05***	***2***	***0.46***	***active***	***Y***
Maneb	12427–38-2	0.00	inactive	1.00	TN	NA	NA		NA	inactive	Y

Overall classifications and potencies of 20 hit expansion substances as compared to the ToxCast ER Agonist Score. The similar substances for the nine (excluding EDTA) potential 'Novel estrogenic substances' (bold italic) were selected based on their structural similarity and their activity in different ToxCast steroidal nuclear receptor assays. EC50, mean potency from multiple independent runs (n). SD, standard deviation. ER Bioactivity, potency (logEC50) normalized to 17-alpha-Ethinylestradiol (1.00). NA, not available/not applicable. FN, false negative substances. TN, true negative substance.

[a] Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015)

[b] CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016)

Table 4
Development and evaluation (cross-validation) of CP ER models.

U.S. EPA *in chemico/in vitro* ER screening assays			Conformal prediction ER models										
Assay name	Assay ID (aeid)	Biological mechanism	Validity			Efficiency			Accuracy			no. of compounds	
			all	inactive	active	all	inactive	active	all	inactive	active	inactive	active
NVS_NR_hER	714	receptor binding	0.85	0.85	0.85	0.40	0.36	0.57	0.73	0.69	0.81	819	204
OT_ER_ERaERb_1440	745	receptor dimerization	0.86	0.85	0.88	0.32	0.29	0.53	0.68	0.62	0.88	1333	180
ATG_ERE_CIS_up	75	gene expression, mRNA	0.85	0.85	0.85	0.43	0.40	0.51	0.74	0.71	0.81	2257	792
ATG_ERa_TRANS_up	117	gene expression, mRNA	0.84	0.84	0.84	0.50	0.47	0.61	0.77	0.75	0.83	2385	681
TOX21_ERa_BLA_Agonist_ratio	785	gene expression, protein	0.86	0.86	0.84	0.57	0.57	0.57	0.87	0.87	0.84	6465	320
TOX21_ERa_LUC_VM7_Agonist	788	gene expression, protein	0.85	0.85	0.85	0.25	0.22	0.41	0.58	*0.5*	0.84	5719	858
ACEA_T47D_80hr_Positive	2	cell proliferation	0.86	0.87	0.84	0.24	0.20	0.41	0.63	*0.54*	0.81	1307	266
		Mean	**0.85**			**0.39**			**0.71**	**0.67**	**0.83**		
		SD	**0.01**			**0.12**			**0.10**	**0.13**	**0.03**		

Information on the seven CP models built for seven ER screening assay endpoints conducted in the ToxCast project. Accuracies highlighted in italics indicate reduced performance of two CP ER models for prediction of inactive substances.

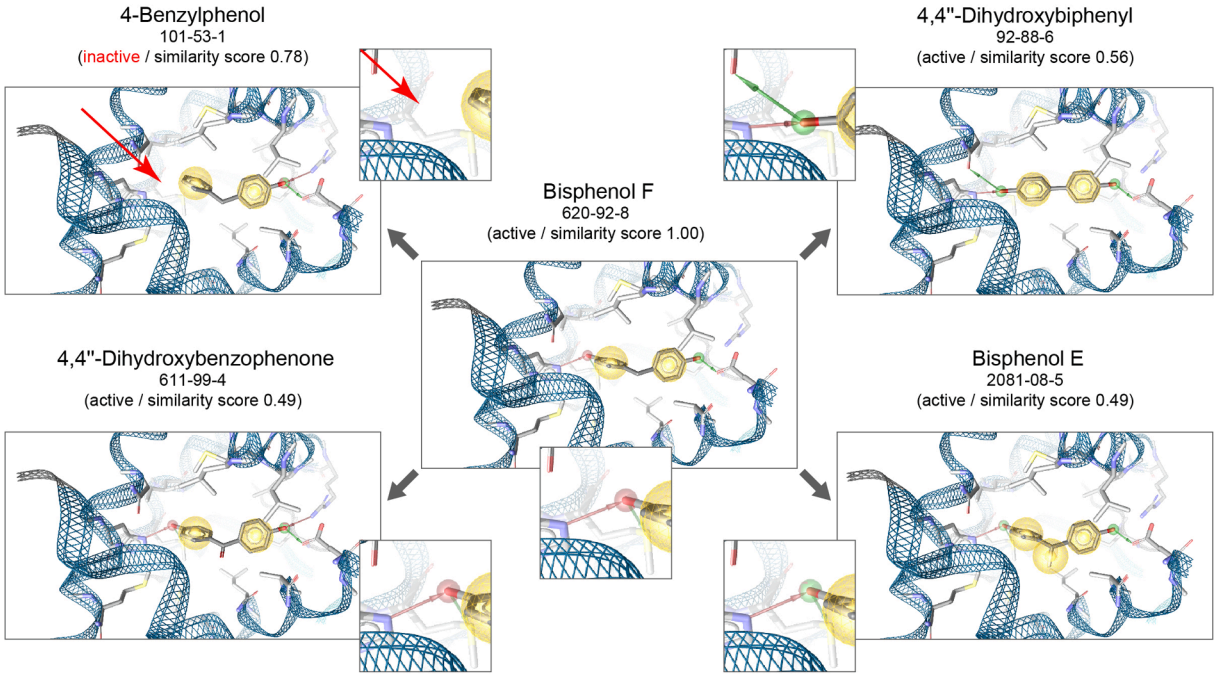

Fig. 7. Docking of bisphenols into ERα. Pharmacophoric interactions of different bisphenols with ERα. For each substance, the outcome of the E-Morph Screening Assay (active/inactive) and the Tanimoto similarity to Bisphenol F are indicated. The red arrow points to a clear difference in the binding of 4-Benzylphenol to ERα (missing hydrogen bond) as compared to the other four tested bisphenols. Visualizations using LigandScout after docking with the @TOME-2 webserver according to (Delfosse et al. 2012). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

substances, no conclusive (Table S4, 'NC') CP ER model classifications were achieved, i.e., 'both class' predictions were returned by the CP framework (no decisions could be made), and not all these substances were tested in every of the seven ER screening assays (Table S4, 'NA').

For this small subset of 29 substances, the overall concordance of the CP ER consensus model, the E-Morph Screening Assay, and the consensus of the ER screening assay test results was in the range of 71–76% (Table 6). Hence, the performance of the CP ER consensus model was comparable to the mean accuracy of the seven individual CP ER models (see Table 4) supporting the consensus model approach, which integrates multiple relevant mechanistic events of estrogen signaling. Importantly, the high predictivity of the CP ER consensus model (100%) and the E-Morph Screening Assay (87%) for active class substances promotes their future use in HTS frameworks. With regard to the evaluation of the E-Morph Assay screening results, the CP ER consensus model supported 89% of the active class assignments for the subset of 29 substances. The reduced predictivity of the E-Morph Screening Assay (54%) for inactive class substances was mainly caused by a higher frequency of non-concordant 'false positive' results, which, however, represent substances with potential estrogenic activity ('Novel estrogenic substances' group, Table S1) that are of particular interest for prioritized follow-up testing.

Taken together, these data suggest that the generated CP ER models are applicable for fast and efficient browsing of large substance libraries to prioritize substances with potential estrogenic activity for subsequent *in vitro* testing in HTS approaches, such as E-Morph. Benchmarking the CP prediction models and the E-Morph Screening Assay against a larger set of reference substances will provide further insights into their

Table 5
Comparison of results from the E-Morph Screen and CP ER models with published *in vitro* and *in silico* ER data from the U.S. EPA.

Chemical name	CAS No.	U.S. EPA *in silico* ER models		U.S. EPA *in chemico/in vitro* ER screening assays		E-Morph Screening Assay		Conformal prediction ER models	
		ToxCast	CERAPP	Consensus **test results**		Substance group	Potency [M]	Consensus **predictions**	
		ER Agonist Score[a)]	ER Agonist Model[b)]	active	inactive		EC50	active	inactive
Hexythiazox	***78587–05-0***	***0.00***	***inactive***	***1 (14%)***	***6 (86%)***	***Novel***	***1.01E-08***	***0***	***0***
Norethindrone acetate (NETA)	***51–98-9***	***NA***	***active***	***2 (100%)***	***0***	***Novel***	***6.25E-07***	***6 (100%)***	***0***
Norgestimate	35189–28-7	NA	active	2 (100%)	0	Novel	7.27E-07	4 (100%)	0
Nandrolone	***434–22-0***	***NA***	***active***	***2 (100%)***	***0***	***Novel***	***2.04E-06***	***7 (100%)***	***0***
Phloretin	***60–82-2***	***NA***	***active***	***2 (100%)***	***0***	***Novel***	***3.36E-06***	***7 (100%)***	***0***
Benzophenone-2	131–55-5	0.40	active	7 (100%)	0	Known	3.55E-06	7 (100%)	0
Troclocarban	101–20-2	0.00	inactive	0	6 (100%)	Novel	3.65E-06	1 (100%)	0
2,4,6-Tri-tert-butylphenol (TTBP)	***732–26-3***	***0.00***	***inactive***	***1 (20%)***	***4 (80%)***	***Novel***	***4.73E-06***	***4 (80%)***	***1 (20%)***
Bisphenol F	***620–92-8***	***NA***	***active***	***4 (100%)***	***0***	***Novel***	***4.79E-06***	***7 (100%)***	***0***
2,4,4′-Trihydroxybenzophenone	1470–79-7	NA	active	5 (100%)	0	Novel	5.60E-06	7 (100%)	0
Diuron	***330–54-1***	***0.00***	***inactive***	***0***	***6 (100%)***	***Novel***	***6.04E-06***	***0***	***5 (100%)***
Fluoxastrobin	361377–29-9	0.00	inactive	1 (20%)	4 (80%)	Novel	6.23E-06	0	0
Azoxystrobin	***131860–33-8***	***0.00***	***inactive***	***1 (17%)***	***5 (83%)***	***Novel***	***6.34E-06***	***1 (100%)***	***0***
4,4′-Dihydroxybenzophenone	611–99-4	NA	active	2 (100%)	0	Novel	1.17E-05	7 (100%)	0
Bisphenol E	2081–08-5	NA	active	4 (100%)	0	Novel	1.34E-05	7 (100%)	0
4,4′-Dihydroxybiphenyl	92–88-6	NA	active	2 (100%)	0	Novel	1.34E-05	2 (100%)	0
Levonorgestrel	797–63-7	0.39	active	6 (86%)	1 (14%)	Known	1.97E-05	6 (100%)	0
Zineb	***12122–67-7***	***NA***	***inactive***	***0***	***0***	***Novel***	***8.71E-05***	***0***	***2 (100%)***
Ethynodiol diacetate	297–76-7	NA	active	2 (100%)	0	Novel	< 1.37E-08	5 (100%)	0
Diphenolic acid	126–00-1	0.17	active	6 (86%)	1 (14%)	Novel	pos ≥ 30 μM	5 (100%)	0
Norgestrel	6533–00-2	0.39	active	5 (100%)	0	FN	inactive	6 (100%)	0
4-Benzylphenol	101–53-1	NA	active	2 (100%)	0	NA	inactive	4 (100%)	0
Picoxystrobin	117428–22-5	0.00	inactive	1 (17%)	5 (83%)	TN	inactive	1 (100%)	0
Butylhydroxytoluene	128–37-0	0.00	inactive	1 (17%)	5 (83%)	TN	inactive	2 (100%)	0
2,5-Di-*tert*-butylhydroquinone	88–58-4	0.00	inactive	1 (14%)	6 (86%)	TN	inactive	4 (100%)	0
Swep	1918–18-9	NA	inactive	0 (0%)	4 (100%)	NA	inactive	0	5 (100%)
Linuron	330–55-2	0.00	inactive	0 (0%)	7 (100%)	TN	inactive	0	1 (100%)
Iprodion	36734–19-7	0.00	inactive	1 (17%)	5 (83%)	TN	inactive	0	0
Maneb	12427–38-2	0.00	inactive	0 (0%)	6 (100%)	TN	inactive	0	2 (100%)

Overall classifications and potencies for the nine (excluding EDTA) potential 'Novel estrogenic substances' (bold italic) and the 20 hit expansion chemicals that were tested in the E-Morph Screening Assay as compared to the ToxCast ER Agonist Score and the available ToxCast ER agonist assay screening data. Consensus predictions considering modes of all single class predictions from the individual ER models or ER screening assays. EC50, mean potency from multiple independent runs. NA, not available/not applicable. FN, false negative substances. TN, true negative substances.
[a)]Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model (Browne et al. 2015).
[b)]CERAPP: Collaborative Estrogen Receptor Activity Prediction Project (Mansouri et al. 2016).

predictive capacities in future studies. In iterative *in silico - in vitro* screening cycles, the newly generated data could then also be used to update and improve the CP ER models. Particularly, additional training data that cover the chemical space for which the current models make poor predictions could prompt more efficient and accurate predictions, which in turn supports the identification of additional active substances by HTS (Svensson et al. 2017a).

4. Conclusion

Over the years, many organizations worldwide have published candidate lists of suspected EDCs, which include hundreds of substances that may pose a potential threat to human health and the environment (WHO/UNEP, 2012). The identification and regulatory restriction of such EDCs is a central goal of chemicals management frameworks and policies worldwide. Commitments such as the 'European Green Deal' even pursue a zero-pollution ambition towards a fully 'toxic-free environment' in the next decades. This intention is reflected in the recently adopted 'Chemicals Strategy for Sustainability' (EC 2020), which also promotes the 'safe-by-design' approach, i.e. the use of substances that pose less or no harm to humans and the environment. For now, adverse health effects of EDCs are mainly investigated in animal experiments (OECD, 2001, 2018a, 2018b, 2018c), although animal data are not necessarily directly translatable to (patho-)physiological processes in humans (Holen et al. 2017). Moreover, these *in vivo* assays are not necessarily specific for individual endocrine mechanisms or suitable for the analysis of 'real-life' co-exposure scenarios. The projected doubling of the global chemical sales by 2030 (WHO, UNEP, 2019) and, in parallel, the intended phasing out of animal experimentation in toxicological testing (Grimm 2019) emphasize the need for novel human-relevant HTS methods and computational approaches to ensure protection of human health and the environment.

Table 6
Predictivity of the E-Morph Screening Assay and CP ER models.

			U.S. EPA *in chemico/in vitro* ER screening assays	E-Morph Screening Assay
			Consensus test results	Test results
Conformal prediction ER models	Consensus predictions	N	25	26
		$N_{True\ Actives}$	15	16
		$N_{False\ Inactives}$	0	2
		$N_{True\ Inactives}$	4	3
		$N_{False\ Actives}$	6	5
		Concordance	**76%**	**73%**
		$P_{active\ class}$	**100%**	**89%**
		$P_{inactive\ class}$	**40%**	**38%**
			U.S. EPA *in chemico/in vitro*ER screening assays	Conformal prediction ER models
			Consensus test results	Consensus predictions
E-Morph Screening Assay	Test results	N	28	26
		$N_{True\ Actives}$	13	16
		$N_{False\ Inactives}$	2	5
		$N_{True\ Inactives}$	7	3
		$N_{False\ Actives}$	6	2
		Concordance	**71%**	**73%**
		$P_{active\ class}$	**87%**	**76%**
		$P_{inactive\ class}$	**54%**	**60%**
			E-Morph Screening Assay	Conformal prediction ER models
			Test results	Consensus test results
U.S. EPA *in chemico/in vitro* ER screening assays	Consensus predictions	N	28	25
		$N_{True\ Actives}$	13	15
		$N_{False\ Inactives}$	6	6
		$N_{True\ Inactives}$	7	4
		$N_{False\ Actives}$	2	0
		Concordance	**71%**	**76%**
		$P_{active\ class}$	**68%**	**71%**
		$P_{inactive\ class}$	**78%**	**100%**

Concordance between the E-Morph Screening Assay, the CP ER models, and the available ToxCast ER agonist assay screening data was calculated based on the results of the nine (excluding EDTA) 'Novel estrogenic substances' and the 20 hit expansion chemicals. Note that the total numbers (n) of substances differ because for some of the 29 test substances, no conclusive CP ER model classifications were achieved and not all of the 29 substances were tested in every of the seven ToxCast ER screening assays. Consensus predictions considering modes of all single class predictions from the individual CP ER models or ToxCast ER screening assays. n, total number of substances. N, number of true/false active/inactive substances for each type of comparison. P, predictivity for active/inactive substances.

4.1. The E-Morph screening Assay provides a reliable and robust human-relevant readout to determine ER signaling activity by phenotypic HTS

The E-Morph Assay (Kornhuber et al. 2021) addresses a human-relevant functional endpoint of adversity, i.e., the perturbation of cell–cell adhesion leading to breast cancer progression and metastasis (Bischoff et al. 2020). In the present study, we further developed the applicability of the original E-Morph Assay for automated HTS using local changes in E-Cad-GFP signal intensity (SI) as a novel, simple and reliable HTS-compatible phenotypic readout for estrogenic activity (see Fig. 1). The SI readout was very robust, with each valid run in the primary screen achieving a Z'-factor above 0.5 (Iversen et al. 2006; Zhang et al. 1999). The determined EC50 values under anti-estrogenic (Fulv treatment) and estrogenic (Fulv + E2 treatment) conditions were directly comparable to the results of the original E-Morph Assay (Kornhuber et al. 2021) with the advantage that the adapted assay avoids both live-cell staining and extensive quantitative image analysis procedures. Based on an intact, complete, and interconnected endogenous estrogen signaling pathway, the E-Morph Screening Assay therefore facilitates the efficient identification of substances with estrogenic activities and the determination of their potencies from concentration-response curves. It could therefore help to accelerate the identification of new substances of concern and support the comprehensive assessment of potential mixture effects of EDCs (Schlotz et al. 2017; Yu et al. 2019).

4.2. E-Morph phenotypic screening correctly identified 27 'known' estrogenic substances and 10 'novel' substances with potential estrogenic activity

We used the E-Morph Screening Assay to analyze a novel substance library (BfR-ChemLibrary) comprising 430 toxicologically-relevant industrial chemicals, biocides, and plant protection products (see Figs. 2 and 3). We identified 27 estrogenic substances of which the potencies of 24 substances correlated very well with the ToxCast ER pathway model (Browne et al. 2015; Judson et al. 2015) (see Fig. 4, Table 1). We further identified 10 additional potential estrogenic substances that have not been described as such in ToxCast before (see Fig. 4, Table 2). According to a recently proposed human-relevant potency threshold (HRPT) for ERα agonism, the minimum relative activity of a test substance must be at least 0.01% of strong estrogens (E2 or 17α-Ethinylestradiol) to exert adverse effects in humans via an ERα-mediated mechanism (Borgert et al. 2018). In the E-Morph Screening Assay, the potencies of the active substances (Fig. 4; Table 1 and 2) were in the range of 1 nM (strong activity, e.g., E2) to 10 μM (weak activity, e.g., Apigenin) and, thus, fitted well into the HRPT with only Zineb displaying a potency slightly below. Subsequent hit verification studies, including gene expression profiling and ERα binding supported the detected estrogenic activity of Hexythiazox, NETA, Nandrolone, Phloretin, Diuron, and Bisphenol F but not of 2,4,6-TTBP, Zineb, and Azoxystrobin (see Fig. 5).

4.3. Use of in silico tools increased the hit-rate and supported the hit evaluation

The E-Morph screening results for the group of 'novel' substances (Table 2) were further substantiated by additional testing of 20 structurally similar substances that were selected based on an *in silico* similarity search for subsequent hit expansion screening, which identified in total another nine ER active substances (see Fig. 6 and Table 3). While being structurally very similar to Bisphenol F, 4-Benzylphenol was inactive in the E-Morph Screening Assay. Additional docking studies detected a difference in the binding mode (i.e., a missing hydrogen bond) to ERα, which could explain the difference in activity of otherwise very similar molecules (see Fig. 7). Hence, computational docking analyses can significantly support the interpretation of *in vitro* screening results, particularly regarding the capability of substances to bind to nuclear hormone receptors. In addition, we built seven *in silico* ER models using the CP framework and the publicly available ToxCast assay data to predict further substances with potential estrogenic activity and to support the E-Morph screening results. The high predictivity for active substances (see Table 5 and 6), which is particularly important from a regulatory point of view to protect human health and the environment, support that the E-Morph Screening Assay and the CP ER models are fit-for-purpose to be applied to new data in an automated manner.

4.4. Future applications of the E-Morph screening Assay and the CP ER models

Provided that a future validation study demonstrates transferability and inter-laboratory reproducibility, the E-Morph Screening Assay appears to be generally suitable for inclusion in existing HTS projects, where it can be used to identify both substances with estrogenic and anti-estrogenic activities using the same phenotypic readout. In ER testing batteries, the E-Morph Screening Assay could be used for efficient analysis of comprehensive substance libraries in order to prioritize substances for subsequent testing against higher tier endpoints, thereby avoiding unnecessary animal testing. ER testing strategies could additionally benefit from further development and implementation of *in silico* tools, including similarity search approaches and CP models, in the evaluation of screening results and targeted selection of candidate substances for follow-up *in vitro* analysis. Well-trained CP ER models may ultimately even replace existing ER HTS assays that resemble the complex (patho-)physiological processes in humans to a rather limited extent. The combination of human-relevant HTS assays and CP models in testing and assessment strategies can ultimately help to increase confidence in *in vitro* results for the regulatory decisions making and thus make an important contribution to achieve the goal of a next generation risk assessment framework that does no longer depend on animal experimentation.

CRediT authorship contribution statement

Saskia Klutzny: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Marja Kornhuber:** Validation, Investigation, Data curation, Writing – review & editing. **Andrea Morger:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Gilbert Schönfelder:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Andrea Volkamer:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Michael Oelgeschläger:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration. **Sebastian Dunst:** Conceptualization, Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper: [The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A patent application (published under EP3517967, WO/2019/145517, US20210055311) for the endpoint and conceptual design of the E-Morph Assay to screen substances for estrogenic or anti-estrogenic activity has been filed at the European Patent Office by the employer (German Federal Institute for Risk Assessment (BfR)) of the authors. The German Federal Institute for Risk Assessment (BfR) is a scientifically independent institution within the portfolio of the Federal Ministry of Food and Agriculture (BMEL) in Germany. The authors' freedom to design, conduct, interpret, and publish research is explicitly not compromised.].

Acknowledgements

We thank Sylvie Coscoy (Laboratoire Physico-Chimie Curie, Institut Curie, PSL Research University - Sorbonne Universités, UPMC-CNRS, Paris, France) for providing the MCF-7/E-Cad-GFP cell line. We are particularly grateful to Verena Fetz (BfR, Berlin, Germany) for supporting the set-up of automated HTS workflows and Edgar Specker and Marc Nazaré (Compound Management Unit and Medicinal Chemistry Group, Leibniz Institute of Molecular Pharmacology, Berlin, Germany) for generation and management of the BfR-ChemLibrary compound plates. We further gratefully acknowledge our colleagues from BfR/Bf3R for scientific input and comments on the manuscript. This work was supported by an internal BfR research funding program (Sonderforschungsprojekt 1322-683), the HaVo-Stiftung for A.M., and a BMBF grant (031A262C) for A.V.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106947.

References

Alvarsson, J., Arvidsson McShane, S., Norinder, U., Spjuth, O., 2021. Predicting With Confidence: Using Conformal Prediction in Drug Discovery. J. Pharm. Sci. 110 (1), 42–49.

Atkinson, F. Standardiser. https://github.com/flatkinson/standardiser; 2014.

Bell, S., Abedini, J., Ceger, P., Chang, X., Cook, B., Karmaus, A.L., Lea, I., Mansouri, K., Phillips, J., McAfee, E., Rai, R., Rooney, J., Sprankle, C., Tandon, A., Allen, D., Casey, W., Kleinstreuer, N., 2020. An integrated chemical environment with tools for chemical safety testing. Toxicol. In Vitro 67, 104916. https://doi.org/10.1016/j.tiv.2020.104916.

Bell, S.M., Phillips, J., Sedykh, A., Tandon, A., Sprankle, C., Morefield, S.Q., Shapiro, A., Allen, D., Shah, R., Maull, E.A., Casey, W.M., Kleinstreuer, N.C., 2017. An Integrated Chemical Environment to Support 21st-Century Toxicology. Environ. Health Perspect 125, 054501.

Bender, A., Glen, R.C., 2004. Molecular similarity: a key technique in molecular informatics. Org. Biomol. Chem. 2 (22), 3204. https://doi.org/10.1039/b409813g.

Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. Data Analysis, Machine Learning and Applications; 2008.

Bischoff, P.; Kornhuber, M.; Dunst, S.; Zell, J.; Fauler, B.; Mielke, T.; Taubenberger, A.V.; Guck, J.; Oelgeschlager, M.; Schonfelder, G. Estrogens Determine Adherens Junction Organization and E-Cadherin Clustering in Breast Cancer Cells via Amphiregulin. iScience 2020;23:101683.

Borgert, C.J., Matthews, J.C., Baker, S.P., 2018. Human-relevant potency threshold (HRPT) for ERalpha agonism. Arch Toxicol 92, 1685–1702.

Branham, W.S.; Dial, S.L.; Moland, C.L.; Hass, B.S.; Blair, R.M.; Fang, H.; Shi, L.; Tong, W.; Perkins, R.G.; Sheehan, D.M. Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor. J Nutr 2002;132:658–664.

Brooijmans, N., Kuntz, I.D., 2003. Molecular recognition and docking algorithms. Annu. Rev. Biophys. Biomol. Struct. 32 (1), 335–373.

Browne, P., Judson, R.S., Casey, W.M., Kleinstreuer, N.C., Thomas, R.S., 2015. Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. Environ. Sci. Technol. 49 (14), 8804–8814.

Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications; 2014.

Carrió, P., Sanz, F., Pastor, M., 2016. Toward a unifying strategy for the structure-based prediction of toxicological endpoints. Arch. Toxicol. 90 (10), 2445–2460.

Chwalisz, K., Surrey, E., Stanczyk, F.Z., 2012. The hormonal profile of norethindrone acetate: rationale for add-back therapy with gonadotropin-releasing hormone agonists in women with endometriosis. Reprod. Sci. 19 (6), 563–571.

de Beco, S., Gueudry, C., Amblard, F., Coscoy, S., 2009. Endocytosis is required for E-cadherin redistribution at mature adherens junctions. Proc. Natl. Acad. Sci. U S A 106 (17), 7010–7015.

de Beco, S.; Gueudry, C.; Amblard, F.; Coscoy, S. Correction for de Beco et al., Endocytosis is required for E-cadherin redistribution at mature adherens junctions. Proc Natl Acad Sci U S A 2020;117:23191.

Delfosse, V., Grimaldi, M., Pons, J.-L., Boulahtouf, A., le Maire, A., Cavailles, V., Labesse, G., Bourguet, W., Balaguer, P., 2012. Structural and mechanistic insights into bisphenols action provide guidelines for risk assessment and discovery of bisphenol A substitutes. Proc. Natl. Acad. Sci. U S A 109 (37), 14930–14935.

Dix, D.J.; Houck, K.A.; Martin, M.T.; Richard, A.M.; Setzer, R.W.; Kavlock, R.J. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol Sci 2007;95:5-12.

EC. European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Chemicals Strategy for Sustainability - Towards a Toxic-Free Environment ed˚eds. Brussels; 2020.

EFSA. The 2015 European Union report on pesticide residues in food. EFSA J 2017;15:e04791.

Filer, D., Patisaul, H.B., Schug, T., Reif, D., Thayer, K., 2014. Test driving ToxCast: endocrine profiling for 1858 chemicals included in phase II. Curr Opin Pharmacol 19, 145–152.

Gayvert, K.M.; Madhukar, N.S.; Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. Cell Chem Biol 2016;23:1294–1301.

Grimm, D., 2019. EPA to eliminate all mammal testing by 2035. Science. https://doi.org/10.1126/science:aaz4593.
Hemmerich, J., Ecker, G.F., 2020. In silico toxicology: From structure-activity relationships towards deep learning and adverse outcome pathways. Wires Comput. Mol. Sci. 10 (4) https://doi.org/10.1002/wcms.v10.410.1002/wcms.1475.
Holen, I., Speirs, V., Morrissey, B., Blyth, K., 2017. In vivo models in breast cancer research: progress, challenges and future directions. Dis. Model. Mech. 10, 359–371.
Huang, R.L.; Xia, M.H. Editorial: Tox21 Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways As Mediated by Exposure to Environmental Toxicants and Drugs. Front Env Sci-Switz 2017;5.
Iversen, P.W., Eastwood, B.J., Sittampalam, G.S., Cox, K.L., 2006. A comparison of assay performance measures in screening assays: signal window, Z' factor, and assay variability ratio. J. Biomol. Screen 11 (3), 247–252.
Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: prediction of molecular toxicity with confidence. Bioinformatics 2018;34:2508–2509.
Jordan, V.C., 1977. Effects of tamoxifen in relation to breast cancer. Br Med J 1 (6075), 1534–1535.
Judson, R.S., Houck, K.A., Kavlock, R.J., Knudsen, T.B., Martin, M.T., Mortensen, H.M., Reif, D.M., Rotroff, D.M., Shah, I., Richard, A.M., Dix, D.J., 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. Environ. Health Perspect. 118 (4), 485–492.
Judson, R.S., Houck, K.A., Watt, E.D., Thomas, R.S., 2017. On selecting a minimal set of in vitro assays to reliably determine estrogen agonist activity. Regul. Toxicol. Pharmacol. 91, 39–49.
Judson, R.S., Magpantay, F.M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M., Huang, R., Rotroff, D.M., Filer, D.L., Houck, K.A., Martin, M.T., Sipes, N., Richard, A.M., Mansouri, K., Setzer, R.W., Knudsen, T.B., Crofton, K.M., Thomas, R. S., 2015. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. Toxicol. Sci. 148 (1), 137–154.
Kim, S., Thiessen, P.A., Bolton, E.E., Bryant, S.H., 2015. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. Available Nucleic Acids Res 43, W605–W611. https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest.
Kleinstreuer, N.C., Ceger, P.C., Allen, D.G., Strickland, J., Chang, X., Hamm, J.T., Casey, W.M., 2016. A Curated Database of Rodent Uterotrophic Bioactivity. Environ Health Perspect 124 (5), 556–562.
Kornhuber, M., Dunst, S., Schönfelder, G., Oelgeschläger, M., 2021. The E-Morph Assay: Identification and characterization of environmental chemicals with estrogenic activity based on quantitative changes in cell-cell contact organization of breast cancer cells. Environ. Int. 149, 106411. https://doi.org/10.1016/j.envint.2021.106411.
Kuhl, H., 2005. Pharmacology of estrogens and progestogens: influence of different routes of administration. Climacteric 8 (sup1), 3–63.
Landrum, G.A. RDKit: Open-source cheminformatics. http://www.rdkit.org; 2006.
Linusson, H. Nonconformist. http://donlnz.github.io/nonconformist/; 2015.
Linusson, H.; Norinder, U.; Boström, H.; Johansson, U.; Löfström, T. On the Calibration of Aggregated Conformal Predictors. Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications. Proceedings of Machine Learning Research: PMLR; 201.
Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25, 402–408.
Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V., 2015. Deep neural nets as a method for quantitative structure-activity relationships. J. Chem. Inf. Model. 55 (2), 263–274.
Maggiora, G., Vogt, M., Stumpfe, D., Bajorath, Jürgen, 2014. Molecular similarity in medicinal chemistry. J. Med. Chem. 57 (8), 3186–3204.
Malo, N., Hanley, J.A., Cerquozzi, S., Pelletier, J., Nadon, R., 2006. Statistical practice in high-throughput screening data analysis. Nat. Biotechnol. 24 (2), 167–175.
Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., Zakharov, A., Worth, A., Richard, A.M., Grulke, C.M., Trisciuzzi, D., Fourches, D., Horvath, D., Benfenati, E., Muratov, E., Wedebye, E.B., Grisoni, F., Mangiatordi, G. F., Incisivo, G.M., Hong, H., Ng, H.W., Tetko, I.V., Balabin, I., Kancherla, J., Shen, J., Burton, J., Nicklaus, M., Cassotti, M., Nikolov, N.G., Nicolotti, O., Andersson, P.L., Zang, Q., Politi, R., Beger, R.D., Todeschini, R., Huang, R., Farag, S., Rosenberg, S.A., Slavov, S., Hu, X., Judson, R.S., 2016. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environ. Health Perspect. 124 (7), 1023–1033.
Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. Front. Env. Sci. Switz 2016;3.
Morger, A., Mathea, M., Achenbach, J.H., Wolf, A., Buesen, R., Schleifer, K.J., Landsiedel, R., Volkamer, A., 2020. KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. J. Cheminform. 12, 24.
Morger, A., Svensson, F., McShane, S.A., Gauraha, N., Norinder, U., Spjuth, O., Volkamer, A., 2021. Assessing the Calibration in Toxicological in Vitro Models with Conformal Prediction. J. Cheminformatics.
Norinder, U., Carlsson, L., Boyer, S., Eklund, M., 2014. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J. Chem. Inf. Model. 54 (6), 1596–1603.
Norinder, U., Rybacka, A., Andersson, P.L., 2016. Conformal prediction to define applicability domain - A case study on predicting ER and AR binding. SAR QSAR Environ. Res. 27 (4), 303–316.
OECD. Test No. 416: Two-Generation Reproduction Toxicity ed^eds: OECD Publishing; 2001.
OECD. Test No. 493: Performance-Based Test Guideline for Human Recombinant Estrogen Receptor (hrER) In Vitro Assays to Detect Chemicals with ER Binding Affinity ed^eds: OECD Publishing; 2015.
OECD. Test No. 455: Performance-Based Test Guideline for Stably Transfected Transactivation In Vitro Assays to Detect Estrogen Receptor Agonists and Antagonists ed^eds: OECD Publishing; 2016.
OECD. New Scoping Document on in vitro and ex vivo Assays for the Identification of Modulators of Thyroid Hormone Signalling ed^eds; 2017.
OECD. Test No. 443: Extended One-Generation Reproductive Toxicity Study ed^eds: OECD Publishing; 2018a.
OECD. Test No. 451: Carcinogenicity Studies ed^eds: OECD Publishing; 2018b.
OECD. Test No. 452: Chronic Toxicity Studies ed^eds: OECD Publishing; 2018c.
Paterni, I., Granchi, C., Minutolo, F., 2017. Risks and benefits related to alimentary exposure to xenoestrogens. Crit. Rev. Food Sci. Nutr. 57 (16), 3384–3404.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12(Oct):2825–2830.
Pons, J.-L., Labesse, G., 2009. @TOME-2: a new pipeline for comparative modeling of protein-ligand complexes. Nucleic Acids Res 37 (Web Server), W485–W491.
Proença, S., Escher, B.I., Fischer, F.C., Fisher, C., Grégoire, Sébastien, Hewitt, N.J., Nicol, B., Paini, A., Kramer, N.I., 2021. Effective exposure of chemicals in in vitro cell systems: A review of chemical distribution models. Toxicol. In Vitro 73, 105133. https://doi.org/10.1016/j.tiv.2021.105133.
Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev. Comput. Mol. Sci. 6 (2), 147–172.
Reif, D.M., Martin, M.T., Tan, S.W., Houck, K.A., Judson, R.S., Richard, A.M., Knudsen, T.B., Dix, D.J., Kavlock, R.J., 2010. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. Environ. Health Perspect 118 (12), 1714–1720.
Rochester, J.R., Bolden, A.L., 2015. Bisphenol S and F: A Systematic Review and Comparison of the Hormonal Activity of Bisphenol A Substitutes. Environ. Health Perspect 123 (7), 643–650.
Rothman, M.S., Carlson, N.E., Xu, M., Wang, C., Swerdloff, R., Lee, P., Goh, V.H.H., Ridgway, E.C., Wierman, M.E., 2011. Reexamination of testosterone, dihydrotestosterone, estradiol and estrone levels across the menstrual cycle and in postmenopausal women measured by liquid chromatography-tandem mass spectrometry. Steroids 76 (1-2), 177–182.
Rotroff, D.M., Dix, D.J., Houck, K.A., Knudsen, T.B., Martin, M.T., McLaurin, K.W., Reif, D.M., Crofton, K.M., Singh, A.V., Xia, M., Huang, R., Judson, R.S., 2013. Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. Environ. Health Perspect 121 (1), 7–14.
Russell, W.M.S., Burch, R.L., 1959. Principles of humane experimental technique ed^eds. Methuen, London.
Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9 (7), 676–682.
Schlotz, N., Kim, G.-J., Jäger, S., Günther, S., Lamy, E., 2017. In vitro observations and in silico predictions of xenoestrogen mixture effects in T47D-based receptor transactivation and proliferation assays. Toxicol. In Vitro 45, 146–157.
Sirianni, R., Capparelli, C., Chimento, A., Panza, S., Catalano, S., Lanzino, M., Pezzi, V., Andò, S., 2012. Nandrolone and stanozolol upregulate aromatase expression and further increase IGF-I-dependent effects on MCF-7 breast cancer cell proliferation. Mol. Cell Endocrinol. 363 (1-2), 100–110.
Sun, J., Carlsson, L., Ahlberg, E., Norinder, U., Engkvist, O., Chen, H., 2017. Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. J. Chem. Inf. Model. 57 (7), 1591–1598.
Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71 (3), 209–249.
Svensson, F., Norinder, U., Bender, A., 2017a. Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction. J. Chem. Inf. Model. 57 (3), 439–444.
Svensson, F., Norinder, U., Bender, A., 2017b. Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. Toxicol Res (Camb) 6 (1), 73–80.
Tropsha, A., 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. Mol. Inform. 29 (6-7), 476–488.
U.S. EPA CCTE. Center for Computational Toxicology and Exposure, ToxCast and Tox21 Data Spreadsheet. [Online]. Available: https://figshare.com/articles/dataset/ToxCast_and_Tox21_Data_Spreadsheet/6062503. Accessed: 2017-06-23.
U.S. NIEHS. Reference Chemical Lists for Test Method Development and Evaluation [Online]. Available: https://ntp.niehs.nih.gov/whatwestudy/niceatm/resources-for-test-method-developers/refchem/index.html. Accessed: 2019-07-17.
Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic Learning in a Random World ed^eds. Springer, Boston, MA.
Welinder, C., Ekblad, L., 2011. Coomassie staining as loading control in Western blot analysis. J Proteome Res 10 (3), 1416–1419.
Wetzel, C., Pifferi, S., Picci, C., Gök, C., Hoffmann, D., Bali, K.K., Lampe, A., Lapatsina, L., Fleischer, R., Smith, E.S.J., Bégay, V., Moroni, M., Estebanez, L., Kühnemund, J., Walcher, J., Specker, E., Neuenschwander, M., von Kries, J.P., Haucke, V., Kuner, R., Poulet, J.F.A., Schmoranzer, J., Poole, K., Lewin, G.R., 2017. Small-molecule inhibition of STOML3 oligomerization reverses pathological mechanical hypersensitivity. Nat. Neurosci. 20 (2), 209–218.

WHO/IPCS. IPCS global assessment of the state-of-the-science of endocrine disruptors. WHO/PCS/EDC/022 2002:35-50.
WHO/UNEP. State of the science of endocrine disrupting chemicals - 2012 edˆeds; 2013.
WHO/UNEP. Global Chemicals Outlook II - From Legacies to Innovative Solutions: Implementing the 2030 Agenda for Sustainable Development - Synthesis Report. 2019.
Wild, C.P.; Weiderpass, E.; Stewart, B.W.; editors. World Cancer Report: Cancer Research for Cancer Prevention edˆeds. Lyon, France: International Agency for Research on Cancer. Available from: http://publications.iarc.fr/586. Licence: CC BY-NC-ND 3.0 IGO; 2020.
Wolber, G., Langer, T., 2005. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. Available J. Chem. Inf. Model. 45, 160–169. https://www.inteligand.com/ligandscout/.
Yager, J.D., Davidson, N.E., 2006. Estrogen carcinogenesis in breast cancer. N. Engl. J. Med. 354 (3), 270–282.
Yu, H., Caldwell, D.J., Suri, R.P., 2019. In vitro estrogenic activity of representative endocrine disrupting chemicals mixtures at environmentally relevant concentrations. Chemosphere 215, 396–403.
Zhang, J., Norinder, U., Svensson, F., 2021. Deep Learning-Based Conformal Prediction of Toxicity. J. Chem. Inf. Model. 61 (6), 2648–2657.
Zhang, J.H.; Chung, T.D.; Oldenburg, K.R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J Biomol Screen 1999;4:67-73.