Published in final edited form as:

Environ Sci Technol. 2020 October 06; 54(19): 12202–12213. doi:10.1021/acs.est.0c03982.

Machine Learning Models for Estrogen Receptor Bioactivity and Endocrine Disruption Prediction

Kimberley M. Zorn¹, Daniel H. Foil¹, Thomas R. Lane¹, Daniel P. Russo², Wendy Hillwalker³, David J. Feifarek³, Frank Jones³, William D. Klaren³, Ashley M. Brinkman³, Sean Ekins^{1,*}

¹Collaborations Pharmaceuticals Inc., 840 Main Campus Drive, Lab 3510, Raleigh, NC, USA

²Center for Computational and Integrative Biology, Rutgers University, Camden, NJ, USA

³Global Product Safety, SC Johnson and Son, Inc., Racine, WI, USA

Abstract

The U.S. Environmental Protection Agency (EPA) periodically releases *in vitro* data across a variety of targets, including the estrogen receptor (ER). In 2015, the EPA used this data to construct mathematical models of ER agonist and antagonist pathways to prioritize chemicals for endocrine disruption testing. However, mathematical models require *in vitro* data prior to predicting estrogenic activity, but machine learning methods are capable of prospective prediction from molecular structure alone. The current study describes the generation and evaluation of Bayesian machine learning models grouped by the EPA's ER agonist pathway model, using multiple data types with proprietary software, Assay Central™. External predictions with three test sets of *in vitro* and *in vivo* reference chemicals with agonist activity classifications were compared to previous mathematical model publications. Training datasets were subjected to additional machine learning algorithms and compared with rank normalized scores of internal five-fold cross-validation statistics. External predictions were found to be comparable or superior to previous studies published by the EPA. When assessing six additional algorithms for the training datasets, Assay Central™ performed similarly at a reduced computational cost. This study demonstrates machine learning can prioritize chemicals for future *in vitro* and *in vivo* testing of ER agonism.

Graphical Abstract

*To whom correspondence should be addressed. sean@collaborationspharma.com, Phone: 215-687-1320.

Competing interests:

S.E. is owner, K.M.Z., D.H.F., and T.R.L. are employees, and D.P.R. is a consultant of Collaborations Pharmaceuticals, Inc. All other authors are SC Johnson and Son, Inc. employees.

Keywords

Bayesian; estrogen receptor; machine learning; endocrine disruption

INTRODUCTION

Endocrine disruption is a major focus of toxicology research due to government initiatives to evaluate environmental risks to the population ¹, and the estrogen receptor (ER) is a primary target of interest. However, as with many biological targets, the downstream effects are difficult to anticipate without expensive, low-throughput *in vitro* and *in vivo* testing ². The U.S. Environmental Protection Agency (EPA) established the Endocrine Disruptor Screening Program in 1999 to evaluate the risk of endocrine disruption from exposure to pesticides and potential contaminants in drinking water, particularly investigating ER agonists. Since then, the EPA has developed a battery of 11 low-throughput *in vitro* and *in vivo* screening assays (Tier 1 ³) addressing multiple mechanisms of endocrine disruption. However, a majority of industrial chemicals remain unclassified as endocrine disruptors and in 2012 the EPA began moving toward more rapid and cost-effective methods, such as high throughput and computational screening methods.

The use of publicly available high-throughput assay data to rapidly profile chemicals for potential toxicity is an area of active research ^{4, 5}. The increasing availability of data and computational processing power has been invaluable to the numerous efforts to develop tools which can anticipate endocrine disrupting chemicals, including the Endocrine Disruptor Knowledge Base ⁶, the OPEn structure-activity/property Relationship App models ^{7, 8}, and the Organization for Economic Cooperation and Development (Q)SAR Toolbox ⁹, in addition to others ¹⁰. High-throughput screening programs, such as the EPA's ToxCast or the interagency Toxicity Testing in the 21st Century (Tox21) programs ^{11, 12} periodically release high quality, curated, high-throughput screening assay data across a wide variety of biological targets and processes relevant to toxicity, including those related to endocrine disruption ¹³. In 2015, the EPA used Tox21/ToxCast data from 16 assays of early and intermediate molecular events to construct an ER agonist pathway ^{12, 14, 15, 16, 17}. These assays target multiple biological processes involved in ER agonistic signaling by measuring receptor binding ^{18, 19}, protein dimerization, transcription factor-DNA interactions, RNA transcript levels ²⁰, reporter protein levels ²¹, and cell proliferation ²². The goal of this work was to predict and rank the likelihood of ER bioactivity. The EPA later accepted these *in silico* results for the 1812 chemicals as alternatives to Tier 1 binding, transactivation, and uterotrophic (UT) assays ²³. Later, the 16 assays of the ER agonist pathway were combined and a minimal four assays were determined to be capable of producing similar confidence in ER agonism predictions, highlighting those common in the best performing models ²⁴. More recently, the 16 agonist pathway assays were assigned burst-flag hit-calls, which designated chemicals as inactive if 50% of the maximal response elicited (AC₅₀) fell below its measured cytotoxicity, so as to eliminate false-positives resulting from non-endpoint related cell death ²⁵. All of these computational approaches from the EPA have great utility, but *in vitro* response data for ER agonist pathway assays are required for the mathematical generation of pathway prediction scores and do not consider chemical structure. More

recently the EPA has also highlighted their OPERA models which included a weighted-KNN approach to predict new compounds^{7, 8}. While caveats exist for use of more advanced computational models, Bayesian machine learning methods have convincingly shown applicability to drug discovery and toxicology by predicting bioactivity from molecular structure alone^{26, 27}. This machine learning algorithm is computationally inexpensive to apply and can generate models quickly on an average desktop computer; furthermore, when using fingerprint descriptors of chemicals this method is robust and straightforward to the user without overtraining²⁸.

The current study describes several groups of machine learning models using either *in vitro* Tox21/ToxCast ER bioactivity and hit-call data²⁹, the area-under-the-curve (AUC) output data of the EPA's ER agonist pathway^{16, 17}, or burst-flag hit-call data incorporating cytotoxicity considerations²⁵. The ultimate goal of this work was to derive predictive machine learning models to prioritize chemicals for future *in vitro* and *in vivo* testing of ER agonism. The performance of all machine learning models was evaluated by internal five-fold cross-validation statistics. The Bayesian models generated in proprietary software, Assay Central™^{30–39}, were tested using external predictions with three test sets of low-throughput *in vitro* and *in vivo* reference chemicals utilized in previous EPA publications^{17, 40}. External predictions were generated across separate models for all 16 assays from each source and then evaluated in summation to produce an overall active or inactive prediction. These predictions were compared to the results reported in previous studies published by the EPA for the mathematical ER agonist pathway model^{24, 25}, as well as from the Collaborative Estrogen Receptor Activity Project (CERAPP) agonist consensus model⁴¹. Several alternative machine learning algorithms were also compared using five-fold cross validation to demonstrate whether there were any benefits to be gained from these versus the Bayesian algorithm.

EXPERIMENTAL SECTION

Datasets

Data for each of the 16 assays used in the ER agonist pathway model from the EPA (Table 1), was retrieved from three sources: 1) invitroDBv3.1 from the ToxCast download site²⁹ referred to herein as “ToxCast2019”, 2) the publicly available download of ER agonist pathway model data¹⁷ referred to herein as “Browne2015”, and 3) a recent publication²⁵ referred to herein as “Nelms2018”.

Models were created using one of four types of data: 1) AUC values from the ER agonist pathway model, 2) AC₅₀ values, 3) hit-call data ToxCast *in vitro* assays, or 4) burst-flag hit-call data from Nelms2018. Model groups consisted of either 16 or four individual Bayesian models of data from corresponding ER assays, with the exception of the two single models built from AUC values at two activity thresholds which were considered separate groups. Continuous AC₅₀ models utilized a calculated threshold that is unique to each dataset and is described in the following section. Publications from the EPA describing the mathematical ER agonist model pathway defines an AUC score ≥ 0.0501 (rounded to 0.1) as active, scores between 0.001 and 0.0501 (rounded to 0.1) as inconclusive, and truncating scores <0.001 to zero to be classified as inactive¹⁷. The threshold of 0.1 was used for models built from AUC

values, as well as a lower threshold of 0.01, as this is described as an acceptable method to limit false positives in a publication for CERAPP⁴¹. Each data source-type pair was considered one of 12 model groups; abbreviations and descriptions of model groups are presented in Table 2.

Data from each source were curated into a single file using a proprietary application called Bleach (Molecular Materials Informatics, Inc. in Montreal, Canada). After downloading invitroDBv3.1 summary files²⁹, CAS numbers from the chemical summary file (Chemical_Summary_190226.csv) were used to curate structures and quality control notes (i.e. water samples, mixtures, ill-defined) from the EPA Chemistry Dashboard⁴² for 9214 substances. Substances that lacked structures or had valid quality control notes were removed along with other chemicals problematic for machine learning (i.e. polymers, ions). The Bleach application was used to combine structures with identifiers and assay data, as well as to standardize structures for machine learning (i.e. removing neutralizing salts, balancing charges). Manual curation steps included removal of compounds greater than 750 MW after salt removal, with the exception of chemicals composed of less than 50 non-hydrogen atoms. These removals were to specifically target pertinent training data related to consumer products like dyes, rather than compounds like antibiotics and large natural products.

Three test sets were compiled based on reference chemicals used in the previous evaluation of the ER pathway model^{2, 17}. These datasets were separated by *in vitro* agonist potency, *in vivo* agonism, and the *in vivo* guideline-like UT calls not included in the previous test set (Table S1). The *in vivo* and *in vitro* test set reference chemical data were downloaded from the NTP Interagency Center for the Evaluation of Alternative Toxicology Methods website⁴³, and UT call data was taken from supporting information of Browne *et al.*¹⁷. It should be noted that additional chemicals annotated in Table S1 were ultimately removed from the machine learning model training datasets due to identical structures. Chemicals present in more than one test set were retained in all and each test set was evaluated separately. Test set chemicals were standardized as previously stated with Bleach. New training models were generated with a proprietary workflow which removes chemicals from each of the three test sets based on structure to generate three new training models within each group. This generation of training datasets allowed a true external prediction of the test sets from chemical structure alone. The prediction accuracy for each test set was compared to the results described by the EPA's agonist pathway model¹⁷ and the CERAPP agonist consensus model⁴¹ available at the time of manuscript preparation through the EPA Chemistry Dashboard⁴². Antagonist compounds with no agonist activity provided (i.e. 4-hydroxytamoxifen, tamoxifen, reserpine, and progesterone in the *in vitro* test set) and chemicals with both positive and negative UT calls (i.e. "EQUIV" in the *in vivo* guideline-like test set) were excluded from these comparisons.

Assay Central™

The Assay Central™ software has been previously described^{30–39}. Structure-activity datasets collated in Molecular Notebook (Molecular Materials Informatics, Inc. in Montreal, Canada) from various sources are thoroughly curated to generate a Bayesian machine

learning model with multiple scripts. By employing a series of rules to detect problematic data, corrections can be implemented by a combination of automated and human re-curation for structure standardization. The produces a high-quality dataset and a Bayesian model to predict activities for proposed compounds. These models utilize extended-connectivity fingerprints of maximum diameter 6 (ECFP6) descriptors generated from the Chemistry Development Kit library⁴⁴. These descriptors are circular topological fingerprints generated by applying the Morgan algorithm that have widely been noted for their ability to map structure-activity relationships²⁸. From all training set molecules, the Assay Central™ software enumerates all fingerprints from the training set and determines a given fingerprint's contribution to a binary activity classification from the ratio of its presence in active and inactive molecules. The summation of these contributions for a given molecule produces the probability-like score; using the standard probability cutoff, a score of 0.5 or greater designates a chemical as active^{28, 45}. For the work herein, an automated method to select the activity threshold was applied to individual models of continuous AC₅₀ data, such that a threshold for a given model is chosen at a point where true positives and true negatives are balanced with the false analogs^{28, 45}.

Predictions were evaluated from the raw score generated within the Assay Central™ framework and using the standard probability cutoff²⁸ of 0.5. Sets of either 16 or four agonist-pathway assays were utilized for predictions: if more than half of the models in a group predicted a chemical as active then it was considered an overall active agonist prediction. The one exception to this majority-rule classification was for the two models created with AUC values at different thresholds; these were stand-alone models and only assigned a single prediction score to a given chemical, but activity was interpreted in the same way. Classifications were then compared to the reported agonism or UT call (Table S1) without potency considerations to understand the conservatism of the models generated (i.e. both “strong” and “weak” agonist reference chemicals were considered active). Each chemical is also assigned an applicability score, which is a measure of the prospective chemical's molecular fingerprints present in the training model: a higher score implies representation in the model, and more confidence in the prediction score. There is no definitive threshold for an acceptable score, but these were utilized to investigate inaccurate predictions.

Each model includes several metrics to evaluate and compare predictive performance as previously described in a relevant publication³⁶, including Receiver Operator Characteristic, Recall, Precision, F1 Score, Cohen's Kappa^{46, 47}, and Matthews Correlation Coefficient⁴⁸. A Domain metric is also provided as a measure of chemical space covered by a model in relation to the ChEMBL 25 database³⁹; a higher value suggests that the training data is more generally applicable and relevant to a prospective chemical.

Comparison of machine learning algorithms

The complete (i.e. full training and testing) datasets were subjected to further comparison with alternative machine learning algorithms, including random forest, k-Nearest Neighbors, support vector classification, naïve Bayesian, AdaBoosted decision trees and deep learning

architecture; these algorithms, as well as parameter optimization have been previously described^{36, 44}.

In order to compare the differences in the five-fold cross validation scores between machine learning algorithms a rank normalized metric was first obtained by range-scaling all metrics for each model to [0, 1] and then designating the mean as the rank normalized score. Previous work from this group^{39, 49} and that of others⁵⁰ has used this rank normalized score as a performance criterion for each machine learning method. This allows for a comprehensive overall model robustness comparison for different machine learning algorithms. The distribution of the rank normalized scores per machine learning algorithm suggests that they are not normally distributed, therefore nonparametric comparisons were done. The rank normalized scores were compared both pairwise (machine learning comparison per training set) or independently (general machine learning comparison). In addition the “difference from the top” (RNS) metric was also assessed, which is the rank normalized score for each machine learning algorithm subtracted from the highest rank normalized score from that particular training set³⁹. This metric maintains the pair-wise results from each training set cross-validation score by algorithm, which allows the direct comparison of the performance of two machine learning algorithms while maintaining information from the other machine learning algorithms tested simultaneously.

RESULTS

Bayesian machine learning models were created from two releases of Tox21/ToxCast hit-calls and continuous AC₅₀ data, the AUC scores output from the EPA’s ER agonist pathway model and burst-flag hit-calls using Assay Central™. The final, complete (i.e. includes test sets) Bayesian machine learning model performance metrics are summarized in Table 3. Those model groups built with AC₅₀ data were the only ones that utilized inconsistent activity thresholds across assays due to the automated selection to optimize model internal performance. In the Tox21/ToxCast data an inactive AC₅₀ is provided as 1,000,000 µM and active chemical AC₅₀ values usually below 200 µM. In most cases the threshold was calculated at a point that covered all active AC₅₀ values, but there were cases where the threshold was considerably lower as reflected by the number of active chemicals (Table 3).

Alternative machine learning algorithms’ five-fold cross-validation metrics are presented in Figure S1. Assay Central™ Bayesian models had comparable performance metrics to other methods generally, but Table S5 and Figure S2 shows that AdaBoosted decision trees and naïve Bayesian algorithms were consistently outperformed by the other methods. In addition, two of the three comparisons (RNS and rank normalized score with pairwise-comparison) show Assay Central™ statistically significantly outperformed both deep learning architecture and k-Nearest Neighbors (Table S5, Figure S2).

A total of three test sets were utilized to evaluate the predictive performance on chemicals outside the training set. New training models were generated by removing the test set chemicals from each of the final, complete models in each group. In this study antagonist reference chemicals that have no ER agonist information were not evaluated as this is outside the scope of this study. Additionally, any chemicals with a reported “EQUIV” UT

call were also not evaluated since there is not a clear classification. Each test set chemical was a prediction score from either all 16 agonist pathway assays^{16, 17} or the minimal four assays²⁴, and a ‘majority rule’ method was applied to yield an overall classification of predicted agonism. These classifications were compared to the agonist pathway model results reported by Browne *et al.*¹⁷ as well as the CERAPP agonist consensus potency level⁴¹ available through the EPA Chemistry Dashboard⁴¹. For consistency of comparison, binary classification of activity was evaluated rather than specific potency (i.e. despite if a chemical was “Active (Weak)” or “Active (Strong)”, it was considered active).

Prediction accuracies for the *in vitro* reference test set of 40 chemicals (28 active and 12 inactive) across Bayesian machine learning model groups are presented in Figure 1 and Table S2A. The best performing groups for the active reference chemicals in this test set were ToxCast2019-HC-16, ToxCast2019-HC-4, and ToxCast2019-AC50-4. In all three cases 26/28 active agonist reference chemicals were correctly predicted, followed closely by Browne2015-HC-16 correctly predicting 25/28. Almost all model groups correctly identified 10/12 inactive compounds from the *in vitro* test set, however there were varying degrees of false negative predictions. The EPA’s ER agonist pathway model was able to correctly identify 25/28 active chemicals and 11/12 inactive chemicals of the 40 *in vitro* agonist reference chemicals (two inconclusive scores)¹⁷. The CERAPP agonist consensus model correctly identified 21/28 active and 9/12 inactive chemicals (three not scored).

Three chemicals from the *in vitro* reference set inaccurately classified as inactive agonists across most model groups were dicofol, diethylhexyl phthalate, and dibutyl phthalate (Table 4, Table S2A). The CERAPP consensus model and ER agonist pathway model had similar issues; CERAPP assigns inactive classifications to all these chemicals while the ER agonist pathway model scored dibutyl phthalate inconclusive and the other two as inactive¹⁷. In contrast to Browne *et al.*,¹⁷ Bayesian model groups were able to correctly and consistently predict haloperidol as inactive, but incorrectly predicted fenarimol and kepone as inactive as well as spironolactone and corticosterone as active chemicals (Table 4, Table S2A). The CERAPP agonist consensus model was similarly inaccurate for all these chemicals with the exception of spironolactone and also classified benzyl butyl phthalate as inactive. All of these *in vitro* active agonist chemicals have reported potencies of ‘weak’ or ‘very weak’.

Prediction accuracies for the *in vivo* reference test set of 43 chemicals (30 active and 13 inactive) across Bayesian model groups are presented in Figure 2 and Table S2B. All groups were able to accurately predict 28/30 active chemicals, but despite most of these model groups being excellent predictors of active agonists from this test set, few were able to predict inactive chemicals accurately. Only the Nelms2018–16 and Nelms2018–4 groups in addition to the Browne2015-AC50-4 group were able to predict 11/13 inactive reference chemicals, while also limiting false negatives. The EPA’s ER agonist pathway model was able to correctly identify 29/30 active chemicals and 8/13 inactive chemicals (four inconclusive scores)¹⁷. The CERAPP agonist consensus model correctly identified 27/30 active and 10/13 inactive chemicals (three not scored).

Chemicals from the *in vivo* reference set that were consistently inaccurately classified by Assay Central™ model groups and CERAPP were two inactive agonists, 4-hydroxybenzoic

acid and kaempferol, and active agonist octamethylcyclotetrasiloxane (Table 4, Table S2B); the latter two were scored inaccurately by the ER agonist pathway model. Browne *et al.*¹⁷ had difficulty accurately classifying the latter chemical and cites that multiple UT studies classifying octamethylcyclotetrasiloxane as active; the authors proposed the high volatility of the chemical may lead to lower experimental concentrations and inactive measurements *in vitro*. Interestingly, dibutyl and diethylhexyl phthalate, as well as kaempferol are inactive *in vivo* reference chemicals but are reported ‘very weak’ agonists in the *in vitro* test set (Table S1). All other *in vivo* active compounds from this test set were predicted accurately across most Bayesian model groups, which makes it difficult to evaluate predictive performances of each model group with respect to this test set.

Prediction accuracies for the *in vivo* guideline-like (evaluated in only a single study or with contradictory results²) test set of 49 chemicals (15 actives and 34 inactives) across machine learning model are presented Figure 3. The AUC-0.01 model was able to correctly predict 12/15 active reference chemicals at the cost of excessive false positive predictions, as it also returned the most active designations (30/49) for this test set. The next best groups, ToxCast20199-HC-16 and ToxCast2019-HC-4, correctly predicted 10/15 active chemicals while comparatively limiting false positives. Most model groups were able to accurately predict at least 24/34 inactive reference chemicals, but the Nelms2018–16 group was best at predicting inactive reference chemicals (29/34 correct). The EPA’s ER agonist pathway model was able to correctly identify 8/15 active chemicals and 24/34 inactive chemicals (8 inconclusive scores)¹⁷. The CERAPP agonist consensus model performed nearly identically, correctly classifying 8/15 active and 25/34 inactive chemicals (four not scored). The agonist pathway model produced the least number of false negatives of all models, but all machine learning and mathematical models evaluated herein generally produced more false positive classifications than the other two test sets.

The *in vivo* guideline-like chemicals were the most difficult to accurately predict with most Bayesian machine learning model groups. This test set was the only one with a majority of inactive reference chemicals, which may have contributed to the inconsistency in predictions compared to the previous two test sets. Five active agonist chemicals from the guideline-like test set were repeatedly predicted inactive by most model groups (Table 4, Table S2C): benzyl salicylate, β -hexachlorocyclohexane, kepone, pendimethalin, and triclosan. Of these, benzyl salicylate was scored inconclusive and triclosan scored inactive, but all others were scored accurately by the ER agonist pathway model; CERAPP correctly predicted benzyl salicylate. Otherwise, only gibberellic acid was a false negative classification by both CERAPP and the agonist pathway model, but was generally predicted correctly by machine learning model groups. Kepone was also unsuccessfully predicted by all Bayesian model groups as well as CERAPP in the *in vitro* reference test set where it is reported as a weak agonist (Table S1–S2). Seven inactive agonists from the *in vivo* “guideline-like” reference chemicals were consistently classified as active by most model groups (Table 4, Table S2C): 2-naphthalenol, 2,4-di-tert-butylphenol, bis(2-ethylhexyl) terephthalate, heptylparaben, oxybenzone, phenolphthalein, and tetrabromobisphenol A. From these chemicals, only 2,4-di-tert-butylphenol, bis(2-ethylhexyl) terephthalate, and tetrabromobisphenol A were correctly classified by both CERAPP and the ER agonist pathway model.

DISCUSSION

Endocrine disruption has become an important area of research as more data arises from environmental chemical exposure to both humans and wildlife ^{2, 51, 52}. Many decades of research have been undertaken by various groups globally ^{10, 53–60}, but arguably some of the most valuable publicly available primary data is generated by governmental organizations like the EPA and European Chemical Agency. Furthermore, the EPA and others are pursuing new methods for evaluating endocrine disruption, as highlighted by the publications on mathematical modeling and prediction of endocrine disruption ^{16, 17, 24, 25, 41}. However, these mathematical models require expensive *in vitro* data from a series of ER assays to generate AUC values for bioactivity determination. Given the thousands of industrial chemicals with little to no toxicological data, screening by generation of empirical data a not feasible ^{16, 41}.

Herein, various machine learning models were built from the same ER assays utilized by the EPA's ER agonist pathway model as well as its AUC output with Assay Central™. These were then evaluated prospective predictions based on chemical structure alone and compared accuracies to the previous work from the EPA ¹⁷. The Bayesian machine learning model groups were able to predict test set chemicals as good as, or better than, the EPA's ER agonist pathway model (Figures 1–3), and predictions were often also consistent with the CERAPP agonist consensus model scores (Table S2). Evaluation of predictions show that some chemicals were consistently inaccurately classified by various model groups (Table 4). Similar trends were seen in the results of the EPA ER models, but two specific examples are discussed below.

Generally, parabens were either classified as active agonists or equivocal ("EQUIV") across test sets; the only exception was heptylparaben which was assigned a negative UT call in the *in vivo* guideline-like test set (Tables S1–3). Most parabens within the *in vivo* guideline-like test set were excluded from accuracy totals due to the "EQUIV" designation, but they are consistently classified as active ER agonists by Bayesian model groups and the CERAPP agonist consensus model (Table S2C). Similarly, tetrabromobisphenol A and phenolphthalein are exceptions to the rule of bisphenol chemicals: other instances of bisphenols in the three test sets are considered active ER agonists (Table S1–2), and all Bayesian model groups predict them accurately. While these chemical classes were incorrectly predicted by most groups, it is consistent with the activity of similar chemicals.

The measured activity in a given assay are frequently inconsistent with overall classification of ER agonism *in vivo*. Phthalates are considered inactive agonists in both guideline-like and reference *in vivo* test sets, but the *in vitro* test set lists the agonist potency as "very weak" for the three phthalate chemicals present. Several bisphenols and parabens within the test sets are reported as inactive in the *in vivo* guideline-like test set but have active AC₅₀ values in several assays (Table S3). Furthermore, ethylparaben has a "EQUIV" UT call but has defined *in vitro* potency, again highlighting the lack of consistency between *in vitro* and *in vivo* activity classifications. It is worth noting that some problematic chemicals discussed here have lower cytotoxicity values, which may contribute to discrepancies in reported activity; such limitations of assay technologies are described in detail by Browne *et al.* ¹⁷.

Previously it has been noted that chemicals from literature with very weak activity are potentially either outside of the testing range of ToxCast/Tox21 assays or false positives ⁴¹.

All machine learning models are only as good as the data that comprise them: if similar chemotypes are consistently considered active in training data, models will predict other structurally similar chemicals as active. Discrepancies between *in vitro* and *in vivo* data will always be a persistent problem for *in silico* predictions, and care should be taken with ensuring both input data and methods of evaluation are comparable across datasets. This misclassification is not an issue of proper representation of these prospective molecules in the training dataset, as evident by the higher applicability scores (Table S4), and similar coverage of chemical space by each assay reflected in the domain scores (Table 3). Instead it is reasonable to conclude that these chemicals are exceptions to general rules picked up by both machine learning and other modeling approaches. Furthermore, all of these chemical classes are quite simplistic, and thus fingerprints are present in many active/inactive training dataset chemicals. Despite these inaccuracies, the similarity of predictions between Bayesian model groups, the ER agonist pathway model, and the CERAPP agonist consensus model indicates that these machine learning results are competitive with other methods currently accepted by the EPA.

There were several other limitations to this study in relation to the work done by the EPA. Most prominently, the ER agonist pathway model allowed for a spectrum of output scores proportional to bioactivities, including an “inconclusive” designation. In comparison, the model groups simply applied a majority rule method to classify probability-like predictions and lacked further insight. Additionally, while the EPA utilized the same 1812 compounds across all Tox21/ToxCast assays covering the same chemical space, the current study curated all the chemicals tested in each assay. While the inclusion of all data is more complete, it also results in inconsistent totals and thus different models cover slightly different chemical property spaces; this could limit the predictive power of external chemicals if there were not enough similarities present in the training data. Assay CentralTM software used an algorithmically calculated activity threshold, and due to the inconsistent distribution of AC₅₀ values that determine a proper threshold, chemicals could be inconsistently classified within individual models. Furthermore, Browne *et al.* ¹⁷ were able to speculate on an *in vivo* mechanism of endocrine disruption and point to the limitations of specific technologies; alternative methods of making similar determinations using machine learning models is a goal of future studies.

While this study has focused primarily on using Assay CentralTM to develop Bayesian models, there is considerable interest in other machine learning approaches. The present and previous studies compared six other machine learning algorithms ³⁶. These previous studies showed the machine learning models perform similarly regardless of dataset or molecular descriptors used. In this study there is certainly variability, but the five-fold cross validation metrics appear generally comparable for most algorithms (Figure S1–S2), highlighting the need for external validation of *in silico* models. In this study only one molecule descriptor type (ECFP6) was used, but future endeavors could include utilizing other descriptors, evaluating external predictions with alternative algorithms and descriptors, as well as generating consensus predictions similar to the CERAPP agonist consensus model.

This work has demonstrated that external test set prediction with the Bayesian models for ER agonists from chemical structure alone has accuracies that are comparable to the published EPA mathematical model, which requires actual *in vitro* data before predictions. Without this testing requirement there is a significant cost saving of time and money by using machine learning models in place of mathematical models. Additionally, there is considerable agreement with the external predictions and the CERAPP agonist consensus model classifications. It is important to note that decades of structure-activity relationships with much smaller ER datasets and limited structural diversity preceded this work^{10, 53–60}, and this study represents a starting point for the incorporation of these advanced machine learning models into the evaluation of endocrine disruption from chemical structure alone. The machine learning methodologies used and compared in this study can also be widely extended to other aspects of endocrine disruption such as the androgen pathway and steroidogenesis. With the continuous addition of newer ER data like ToxCast/Tox21 into the public domain, it is likely that these machine learning models will be further tested, expanded, updated and improved over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Grant information

We kindly acknowledge SC Johnson and Son, Inc. for funding this work. We also acknowledge for NIH funding to develop the software from 1R43GM122196-01 and R44GM122196-02A1 “Centralized assay datasets for modelling support of small drug discovery organizations” from NIGMS and NIEHS for 1R43ES031038-01 “MegaTox for analyzing and visualizing data across different screening systems”. “Research reported in this publication was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Number R43ES031038. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.” We are also grateful to the EPA for providing the Tox21/ToxCast datasets and for feedback on our earlier modeling efforts. We kindly acknowledge Dr. Alex M. Clark (Molecular Materials Informatics, Inc.) for Assay Central™ support and Dr. William C. Kershaw (SC Johnson and Son, Inc.) for constructive criticism.

ABBREVIATIONS USED

AC ₅₀	50% maximal response
AUC	area under the curve
EPA	U.S. Environmental Protection Agency
ER	estrogen receptor
UT	uterotrophic

REFERENCES

1. Shanle EK; Xu W, Endocrine disrupting chemicals targeting estrogen receptor signaling: identification and mechanisms of action. *Chem Res Toxicol* 2011, 24, (1), 6–19. [PubMed: 21053929]

2. Kleinstreuer NC; Ceger PC; Allen DG; Strickland J; Chang X; Hamm JT; Casey WM, A Curated Database of Rodent Uterotrophic Bioactivity. *Environ Health Perspect* 2016, 124, (5), 556–62. [PubMed: 26431337]
3. EPA Endocrine Disruptor Screening Program Tier 1 Battery of Assays. <https://www.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-tier-1-battery-assays>
4. Russo DP; Strickland J; Karmaus AL; Wang W; Shende S; Hartung T; Aleksunes LM; Zhu H, Nonanimal Models for Acute Toxicity Evaluations: Applying Data-Driven Profiling and Read-Across. *Environ Health Perspect* 2019, 127, (4), 47001. [PubMed: 30933541]
5. Kim MT; Huang R; Sedykh A; Wang W; Xia M; Zhu H, Mechanism Profiling of Hepatotoxicity Caused by Oxidative Stress Using Antioxidant Response Element Reporter Gene Assay Models and Big Data. *Environ Health Perspect* 2016, 124, (5), 634–41. [PubMed: 26383846]
6. Ding D; Xu L; Fang H; Hong H; Perkins R; Harris S; Bearden ED; Shi L; Tong W, The EDKB: an established knowledge base for endocrine disrupting chemicals. *BMC Bioinformatics* 2010, 11 Suppl 6, S5.
7. Mansouri K; Grulke CM; Judson RS; Williams AJ, OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminform* 2018, 10, (1), 10. [PubMed: 29520515]
8. Mansouri K OPERA - source code. <https://github.com/kmansouri/OPERA>
9. Anon QSAR Toolbox. <https://qsartoolbox.org>
10. Zhang L; Sedykh A; Tripathi A; Zhu H; Afantitis A; Mouchlis VD; Melagraki G; Rusyn I; Tropsha A, Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol Appl Pharmacol* 2013, 272, (1), 67–76. [PubMed: 23707773]
11. Sun H; Xia M; Austin CP; Huang R, Paradigm shift in toxicity testing and modeling. *AAPS J* 2012, 14, (3), 473–80. [PubMed: 22528508]
12. Dix DJ; Houck KA; Martin MT; Richard AM; Setzer RW; Kavlock RJ, The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* 2007, 95, (1), 5–12. [PubMed: 16963515]
13. Rotroff DM; Dix DJ; Houck KA; Knudsen TB; Martin MT; McLaurin KW; Reif DM; Crofton KM; Singh AV; Xia M; Huang R; Judson RS, Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ Health Perspect* 2013, 121, (1), 7–14. [PubMed: 23052129]
14. Judson RS; Houck KA; Kavlock RJ; Knudsen TB; Martin MT; Mortensen HM; Reif DM; Rotroff DM; Shah I; Richard AM; Dix DJ, In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect* 2010, 118, (4), 485–92. [PubMed: 20368123]
15. EPA, U. S., EPA's ToxCast Program. 2019.
16. Judson RS; Magpantay FM; Chickarmane V; Haskell C; Tania N; Taylor J; Xia M; Huang R; Rotroff DM; Filer DL; Houck KA; Martin MT; Sipes N; Richard AM; Mansouri K; Setzer RW; Knudsen TB; Crofton KM; Thomas RS, Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 In Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol Sci* 2015, 148, (1), 137–54. [PubMed: 26272952]
17. Browne P; Judson RS; Casey WM; Kleinstreuer NC; Thomas RS, Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. *Environ Sci Technol* 2015, 49, (14), 8804–14. [PubMed: 26066997]
18. Knudsen TB; Houck KA; Sipes NS; Singh AV; Judson RS; Martin MT; Weissman A; Kleinstreuer NC; Mortensen HM; Reif DM; Rabinowitz JR; Setzer RW; Richard AM; Dix DJ; Kavlock RJ, Activity profiles of 309 ToxCast chemicals evaluated across 292 biochemical targets. *Toxicology* 2011, 282, (1–2), 1–15. [PubMed: 21251949]
19. Sipes NS; Martin MT; Kothiya P; Reif DM; Judson RS; Richard AM; Houck KA; Dix DJ; Kavlock RJ; Knudsen TB, Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays. *Chem Res Toxicol* 2013, 26, (6), 878–95. [PubMed: 23611293]
20. Martin MT; Dix DJ; Judson RS; Kavlock RJ; Reif DM; Richard AM; Rotroff DM; Romanov S; Medvedev A; Poltoratskaya N; Gambarian M; Moeser M; Makarov SS; Houck KA, Impact of

- environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. *Chem Res Toxicol* 2010, 23, (3), 578–90. [PubMed: 20143881]
21. Huang R; Sakamuru S; Martin MT; Reif DM; Judson RS; Houck KA; Casey W; Hsieh JH; Shockley KR; Ceger P; Fostel J; Witt KL; Tong W; Rotroff DM; Zhao T; Shinn P; Simeonov A; Dix DJ; Austin CP; Kavlock RJ; Tice RR; Xia M, Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci Rep* 2014, 4, 5664. [PubMed: 25012808]
22. Rotroff DM; Dix DJ; Houck KA; Kavlock RJ; Knudsen TB; Martin MT; Reif DM; Richard AM; Sipes NS; Abassi YA; Jin C; Stampfl M; Judson RS, Real-time growth kinetics measuring hormone mimicry for ToxCast chemicals in T-47D human ductal carcinoma cells. *Chem Res Toxicol* 2013, 26, (7), 1097–107. [PubMed: 23682706]
23. EPA, U. S. Use of High Throughput Assays and Computational Tools: Endocrine Disruptor Screening Program; Notice of Availability and Opportunity for Comment, 80 Fed. Reg. 118 <https://www.federalregister.gov/articles/2015/06/19/2015-15182/use-of-high-throughput-assays-and-computational-tools-endocrine-disruptor-screening-program-notice>
24. Judson RS; Houck KA; Watt ED; Thomas RS, On selecting a minimal set of in vitro assays to reliably determine estrogen agonist activity. *Regul Toxicol Pharmacol* 2017, 91, 39–49. [PubMed: 28993267]
25. Nelms MD; Mellor CL; Enoch SJ; Judson RS; Patlewicz G; Richard AM; Madden JM; Cronin MTD; Edwards SW, A mechanistic framework for integrating chemical structure and high-throughput screening results to improve toxicity predictions. *Comp Toxicol* 2018, 8, 1–12.
26. Ekins S, Progress in computational toxicology. *J Pharmacol Toxicol Methods* 2014, 69, (2), 115–40. [PubMed: 24361690]
27. Bender A, Bayesian methods in virtual screening and chemical biology. *Methods Mol Biol* 2011, 672, 175–96. [PubMed: 20838969]
28. Clark AM; Dole K; Coulon-Spektor A; McNutt A; Grass G; Freundlich JS; Reynolds RC; Ekins S, Open Source Bayesian Models. 1. Application to ADME/Tox and Drug Discovery Datasets. *J Chem Inf Model* 2015, 55, (6), 1231–45. [PubMed: 25994950]
29. EPA ToxCast & Tox21 Summary Files from invitrodb_v3.1. <https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data> (April 16, 2019),
30. Anantpadma M; Lane T; Zorn KM; Lingerfelt MA; Clark AM; Freundlich JS; Davey RA; Madrid PB; Ekins S, Ebola Virus Bayesian Machine Learning Models Enable New in Vitro Leads. *ACS Omega* 2019, 4, (1), 2353–2361. [PubMed: 30729228]
31. Dalecki AG; Zorn KM; Clark AM; Ekins S; Narmore WT; Tower N; Rasmussen L; Bostwick R; Kutsch O; Wolschendorf F, High-throughput screening and Bayesian machine learning for copper-dependent inhibitors of *Staphylococcus aureus*. *Metallomics* 2019, 11, (3), 696–706. [PubMed: 30839007]
32. Ekins S; Gerlach J; Zorn KM; Antonio BM; Lin Z; Gerlach A, Repurposing Approved Drugs as Inhibitors of Kv7.1 and Nav1.8 to Treat Pitt Hopkins Syndrome. *Pharm Res* 2019, 36, (9), 137. [PubMed: 31332533]
33. Ekins S; Puhl AC; Zorn KM; Lane TR; Russo DP; Klein JJ; Hickey AJ; Clark AM, Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater* 2019, 18, (5), 435–441. [PubMed: 31000803]
34. Hernandez HW; Soeung M; Zorn KM; Ashoura N; Mottin M; Andrade CH; Caffrey CR; de Siqueira-Neto JL; Ekins S, High Throughput and Computational Repurposing for Neglected Diseases. *Pharm Res* 2018, 36, (2), 27. [PubMed: 30560386]
35. Lane T; Russo DP; Zorn KM; Clark AM; Korotcov A; Tkachenko V; Reynolds RC; Perryman AL; Freundlich JS; Ekins S, Comparing and Validating Machine Learning Models for Mycobacterium tuberculosis Drug Discovery. *Mol Pharm* 2018, 15, (10), 4346–4360. [PubMed: 29672063]
36. Russo DP; Zorn KM; Clark AM; Zhu H; Ekins S, Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Mol Pharm* 2018, 15, (10), 4361–4370. [PubMed: 30114914]

37. Sandoval PJ; Zorn KM; Clark AM; Ekins S; Wright SH, Assessment of Substrate-Dependent Ligand Interactions at the Organic Cation Transporter OCT2 Using Six Model Substrates. *Mol Pharmacol* 2018, 94, (3), 1057–1068. [PubMed: 29884691]
38. Wang PF; Neiner A; Lane TR; Zorn KM; Ekins S; Kharasch ED, Halogen Substitution Influences Ketamine Metabolism by Cytochrome P450 2B6: In Vitro and Computational Approaches. *Mol Pharm* 2019, 16, (2), 898–906. [PubMed: 30589555]
39. Zorn KM; Lane TR; Russo DP; Clark AM; Makarov V; Ekins S, Multiple Machine Learning Comparisons of HIV Cell-based and Reverse Transcriptase Data Sets. *Mol Pharm* 2019, 16, (4), 1620–1632. [PubMed: 30779585]
40. Program NT NICEATM Reference Chemical Lists for Test Method Evaluations. <https://ntp.niehs.nih.gov/pubhealth/evalatm/resources-for-test-method-developers/refchem/index.html> (02/21/2019),
41. Mansouri K; Abdelaziz A; Rybacka A; Roncaglioni A; Tropsha A; Varnek A; Zakharov A; Worth A; Richard AM; Grulke CM; Trisciuzzi D; Fourches D; Horvath D; Benfenati E; Muratov E; Wedebye EB; Grisoni F; Mangiatordi GF; Incisivo GM; Hong H; Ng HW; Tetko IV; Balabin I; Kancherla J; Shen J; Burton J; Nicklaus M; Cassotti M; Nikolov NG; Nicolotti O; Andersson PL; Zang Q; Politi R; Beger RD; Todeschini R; Huang R; Farag S; Rosenberg SA; Slavov S; Hu X; Judson RS, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ Health Perspect* 2016, 124, (7), 1023–33. [PubMed: 26908244]
42. Williams AJ; Grulke CM; Edwards J; McEachran AD; Mansouri K; Baker NC; Patlewicz G; Shah I; Wambaugh JF; Judson RS; Richard AM, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* 2017, 9, (1), 61. [PubMed: 29185060]
43. NTP NICEATM Reference Chemical Lists for Test Method Evaluations. <https://ntp.niehs.nih.gov/whatwestudy/niceatm/resources-for-test-method-developers/refchem/index.html>
44. Willighagen EL; Mayfield JW; Alvarsson J; Berg A; Carlsson L; Jeliaskova N; Kuhn S; Pluskal T; Rojas-Cherto M; Spjuth O; Torrance G; Evelo CT; Guha R; Steinbeck C, The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform* 2017, 9, (1), 33. [PubMed: 29086040]
45. Clark AM; Ekins S, Open Source Bayesian Models. 2. Mining a “Big Dataset” To Create and Validate Models with ChEMBL. *J Chem Inf Model* 2015, 55, (6), 1246–60. [PubMed: 25995041]
46. Carletta J, Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 1996, 22, 249–254.
47. Cohen J, A coefficient of agreement for nominal scales. *Education and Psychological Measurement* 1960, 20, 37–46.
48. Matthews BW, Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975, 405, (2), 442–51. [PubMed: 1180967]
49. Korotcov A; Tkachenko V; Russo DP; Ekins S, Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol Pharm* 2017, 14, (12), 4462–4475. [PubMed: 29096442]
50. Caruana R; Niculescu-Mizil A, An empirical comparison of supervised learning algorithms. In 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
51. Takemura H; Sakakibara H; Yamazaki S; Shimoi K, Breast cancer and flavonoids - a role in prevention. *Curr Pharm Des* 2013, 19, (34), 6125–32. [PubMed: 23448447]
52. Rodgers KM; Udesky JO; Rudel RA; Brody JG, Environmental chemicals and breast cancer: An updated review of epidemiological literature informed by biological mechanisms. *Environ Res* 2018, 160, 152–182. [PubMed: 28987728]
53. Waller CL; Oprea TI; Chae K; Park HK; Korach KS; Laws SC; Wiese TE; Kelce WR; Gray LE Jr., Ligand-based identification of environmental estrogens. *Chem Res Toxicol* 1996, 9, (8), 1240–8. [PubMed: 8951225]
54. Waller CL, A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *J Chem Inf Comput Sci* 2004, 44, (2), 758–65. [PubMed: 15032558]

55. Waller CL; Minor DL; McKinney JD, Using three-dimensional quantitative structure-activity relationships to examine estrogen receptor binding affinities of polychlorinated hydroxybiphenyls. *Environ Health Perspect* 1995, 103, (7–8), 702–7. [PubMed: 7588482]
56. Asikainen AH; Ruuskanen J; Tuppurainen KA, Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ Res* 2004, 15, (1), 19–32. [PubMed: 15113066]
57. Suzuki T; Ide K; Ishida M; Shapiro S, Classification of environmental estrogens by physicochemical properties using principal component analysis and hierarchical cluster analysis. *J Chem Inf Comput Sci* 2001, 41, (3), 718–26. [PubMed: 11410051]
58. Sakkiiah S; Selvaraj C; Gong P; Zhang C; Tong W; Hong H, Development of estrogen receptor beta binding prediction model using large sets of chemicals. *Oncotarget* 2017, 8, (54), 92989–93000. [PubMed: 29190972]
59. Niu AQ; Xie LJ; Wang H; Zhu B; Wang SQ, Prediction of selective estrogen receptor beta agonist using open data and machine learning approach. *Drug Des Devel Ther* 2016, 10, 2323–31.
60. Bhatarai B; Wilson DM; Price PS; Marty S; Parks AK; Carney E, Evaluation of OASIS QSAR Models Using ToxCast in Vitro Estrogen and Androgen Receptor Binding Data and Application in an Integrated Endocrine Screening Approach. *Environ Health Perspect* 2016, 124, (9), 1453–61. [PubMed: 27152837]

Figure 1:

Accuracy of predictions for the *in vitro* reference chemical test set across all Bayesian machine learning model groups, in comparison to Browne *et al.*¹⁷ as well as CERAPP classifications⁴¹ (collected Dec 5, 2019). Navy bars indicate number of chemicals classified as active by the model group, blue bars indicate the number of correctly classified active chemicals, red bars indicate the number of chemicals classified as inactive by the model group, orange bars indicate the number of correctly classified inactive chemicals. A total of 40 chemicals are included, 28 actives and 12 inactives.

Figure 2:

Accuracy of predictions for the *in vivo* reference chemical test set across all Bayesian machine learning model groups, in comparison to Browne *et al.*¹⁷ as well as CERAPP classifications⁴¹ (collected Dec 5, 2019). Navy bars indicate number of chemicals classified as active by the model group, blue bars indicate the number of correctly classified active chemicals, red bars indicate the number of chemicals classified as inactive by the model group, orange bars indicate the number of correctly classified inactive chemicals. A total of 43 chemicals are included, 30 actives and 13 inactives.

Figure 3:

Accuracy of predictions for the *in vivo* guideline-like test set across all Bayesian machine learning model groups, in comparison to Browne *et al.*¹⁷ as well as CERAPP classifications⁴¹ (collected Dec 5, 2019). Navy bars indicate number of chemicals classified as active by the model group, blue bars indicate the number of correctly classified active chemicals, red bars indicate the number of chemicals classified as inactive by the model group, orange bars indicate the number of correctly classified inactive chemicals. A total of 49 chemicals are included, 15 actives and 34 inactives.

Table 1:

List of assays used for machine learning models (available in ToxCast/Tox21).

Assay Abbreviation	Assay Name
A1	NVS_NR_bER
A2	NVS_NR_hER
A3	NVS_NR_mERa
A4	OT_ER_ERaERa_0480
A5	OT_ER_ERaERa_1440
A6	OT_ER_ERaERb_0480
A7	OT_ER_ERaERb_1440
A8	OT_ER_ERbERb_0480
A9	OT_ER_ERbERb_1440
A10	OT_ERa_EREGFP_0480
A11	OT_ERa_EREGFP_0120
A12	ATG_ERa_TRANS_up
A13	ATG_ERE_CIS_up
A14	TOX21_ERa_BLA_Agonist_ratio
A15	TOX21_ERa_LUC_BG1_Agonist
A16	ACEA_T47D_80hr_Positive

Table 2:

Details of machine learning model groups including abbreviations, number of models, and any associated publications of the data.

No. Models	Data Source	Data Type	Group Abbreviation	Reference	Active Threshold
16	ToxCast/Tox21 (2019 release)	hit-call	ToxCast2019-HC-16		binary
4	ToxCast/Tox21 (2019 release)	hit-call	ToxCast2019-HC-4		binary
16	ToxCast/Tox21 (2019 release)	AC ₅₀	ToxCast2019-AC50-16		automated
4	ToxCast/Tox21 (2019 release)	AC ₅₀	ToxCast2019-AC50-4		automated
16	ToxCast/Tox21 (2014 release)	hit-call	Browne2015-HC-16	16, 17	binary
4	ToxCast/Tox21 (2014 release)	hit-call	Browne2015-HC-4	16, 17	binary
16	ToxCast/Tox21 (2014 release)	AC ₅₀	Browne2015-AC50-16	16, 17	automated
4	ToxCast/Tox21 (2014 release)	AC ₅₀	Browne2015-AC50-4	16, 17	automated
16	ToxCast/Tox21 (2014 modified release)	burst-flag hit-call	Nelms2018-16	25	binary
4	ToxCast/Tox21 (2014 modified release)	burst-flag hit-call	Nelms2018-4	25	binary
1	EPA ER agonist pathway	AUC	AUC-0.1	16, 17	AUC 0.1
1	EPA ER agonist pathway	AUC	AUC-0.01	16, 17	AUC 0.01

Table 3:

Five-fold cross-validation results from final machine learning models. Abbreviations: Receiver Operating Characteristic (ROC), Cohen's Kappa (CK) and Matthews Correlation Coefficient (MCC).

A) ToxCast2019-HC models							
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain
A1	123	1024	0.774	0.495	0.418	0.421	0.283
A2	224	1111	0.746	0.516	0.381	0.383	0.284
A3	181	897	0.720	0.462	0.273	0.296	0.280
A4	146	1732	0.836	0.411	0.330	0.382	0.300
A5	163	1732	0.764	0.418	0.344	0.353	0.300
A6	237	1732	0.854	0.500	0.385	0.431	0.300
A7	314	1732	0.772	0.507	0.375	0.384	0.300
A8	225	1732	0.840	0.473	0.357	0.404	0.300
A9	243	1732	0.766	0.464	0.351	0.367	0.300
A10	185	1732	0.732	0.417	0.336	0.340	0.300
A11	188	1732	0.771	0.368	0.253	0.293	0.300
A12	756	3401	0.779	0.553	0.381	0.407	0.309
A13	856	3401	0.720	0.530	0.361	0.362	0.309
A14	422	7321	0.770	0.315	0.251	0.295	0.369
A15	1169	7321	0.698	0.409	0.267	0.277	0.369
A16	398	2844	0.656	0.357	0.277	0.286	0.297
B) Nelms2018 models							
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain
A1	48	2050	0.883	0.413	0.391	0.455	0.304
A2	104	2712	0.861	0.372	0.333	0.396	0.306
A3	73	2050	0.842	0.316	0.276	0.336	0.304
A4	52	1730	0.836	0.251	0.211	0.303	0.300
A5	61	1730	0.778	0.383	0.354	0.369	0.300
A6	96	1730	0.880	0.335	0.270	0.360	0.300
A7	84	1730	0.788	0.354	0.308	0.333	0.300
A8	94	1730	0.889	0.340	0.277	0.370	0.300
A9	97	1730	0.793	0.430	0.389	0.398	0.300
A10	72	1730	0.820	0.214	0.154	0.241	0.300
A11	67	1730	0.823	0.285	0.239	0.294	0.300
A12	232	3307	0.756	0.307	0.224	0.277	0.309
A13	309	3307	0.711	0.288	0.172	0.210	0.309
A14	171	7309	0.796	0.243	0.214	0.275	0.369
A15	701	7309	0.677	0.279	0.165	0.190	0.369
A16	146	1690	0.724	0.401	0.342	0.343	0.296
C) Browne2015-HC models							

A) ToxCast2019-HC models								
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain	
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain	
A1	91	684	0.793	0.509	0.418	0.427	0.267	
A2	156	765	0.753	0.583	0.480	0.481	0.271	
A3	147	582	0.689	0.514	0.354	0.354	0.265	
A4	133	1681	0.867	0.427	0.353	0.408	0.294	
A5	103	1678	0.818	0.408	0.355	0.381	0.294	
A6	214	1683	0.862	0.504	0.399	0.446	0.294	
A7	194	1677	0.757	0.396	0.290	0.308	0.294	
A8	207	1681	0.832	0.461	0.349	0.397	0.294	
A9	175	1677	0.766	0.438	0.359	0.366	0.294	
A10	142	1684	0.759	0.407	0.341	0.349	0.295	
A11	175	1686	0.785	0.390	0.287	0.322	0.295	
A12	453	1686	0.758	0.580	0.388	0.401	0.295	
A13	529	1686	0.701	0.552	0.336	0.337	0.295	
A14	106	1686	0.788	0.338	0.269	0.323	0.295	
A15	275	1686	0.692	0.439	0.323	0.323	0.295	
A16	306	1682	0.685	0.401	0.283	0.284	0.295	
D) Browne2015-AC50 models								
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain	Activity Threshold
A1	19	1686	0.920	0.343	0.331	0.426	0.295	0.1703 μM
A2	62	1686	0.915	0.420	0.387	0.449	0.295	4.761 μM
A3	94	1686	0.800	0.304	0.238	0.303	0.295	> 1000 μM
A4	133	1686	0.872	0.430	0.356	0.411	0.295	> 1000 μM
A5	103	1686	0.817	0.398	0.343	0.373	0.295	> 1000 μM
A6	214	1686	0.861	0.499	0.393	0.443	0.295	> 1000 μM
A7	107	1686	0.793	0.429	0.378	0.396	0.295	21.69 μM
A8	150	1686	0.873	0.517	0.451	0.489	0.295	42.77 μM
A9	113	1686	0.810	0.381	0.315	0.356	0.295	18.88 μM
A10	76	1686	0.866	0.373	0.329	0.379	0.295	16.53 μM
A11	101	1686	0.849	0.391	0.333	0.385	0.295	19.31 μM
A12	380	1686	0.779	0.567	0.416	0.425	0.295	57.20 μM
A13	104	1686	0.825	0.381	0.320	0.368	0.295	2.774 μM
A14	40	1686	0.884	0.218	0.185	0.279	0.295	20.64 μM
A15	104	1686	0.804	0.305	0.230	0.297	0.295	24.45 μM
A16	132	1686	0.785	0.334	0.246	0.295	0.295	> 1000 μM
E) ToxCast2019-AC50 models								
assay	no. actives	no. total	ROC	F1-score	CK	MCC	domain	activity threshold
A1	123	2057	0.786	0.323	0.256	0.308	0.305	> 1000 μM

A) ToxCast2019-HC models								
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain	
A2	222	2717	0.767	0.443	0.387	0.390	0.307	> 1000 μ M
A3	180	2052	0.762	0.363	0.273	0.310	0.304	> 1000 μ M
A4	145	1732	0.834	0.427	0.352	0.394	0.300	> 1000 μ M
A5	163	1732	0.764	0.418	0.344	0.353	0.300	> 1000 μ M
A6	237	1732	0.854	0.500	0.385	0.431	0.300	> 1000 μ M
A7	273	1732	0.766	0.454	0.313	0.339	0.300	62.48 μ M
A8	165	1732	0.840	0.523	0.459	0.477	0.300	42.77 μ M
A9	127	1732	0.800	0.415	0.349	0.385	0.300	19.66 μ M
A10	110	1732	0.840	0.348	0.279	0.336	0.300	19.31 μ M
A11	81	1732	0.869	0.363	0.315	0.374	0.300	14.34 μ M
A12	286	3403	0.813	0.377	0.289	0.344	0.309	12.22 μ M
A13	382	3403	0.782	0.424	0.322	0.351	0.309	16.22 μ M
A14	293	7321	0.780	0.240	0.187	0.251	0.369	40.78 μ M
A15	556	7321	0.754	0.332	0.251	0.285	0.369	21.38 μ M
A16	397	2844	0.663	0.360	0.280	0.287	0.297	> 1000 μ M
F) AUC models								
Assay	No. Actives	No. Total	ROC	F1-Score	CK	MCC	Domain	Activity Threshold
N/A	262	1686	0.766	0.450	0.313	0.335	0.295	0.01
N/A	87	1686	0.905	0.396	0.344	0.418	0.295	0.1

Table 4:

Chemicals that were inaccurately predicted and the reported activity from each test set.

Structure	Name (CASRN)	Reported Agonist Activity
Diethylhexyl phthalate (117-81-7)	Very Weak Agonist (Vitro) Inactive Agonist (Vivo) Negative UT Call (VivoGL)
Dibutyl phthalate (84-74-2)	Very Weak Agonist (Vitro) Inactive Agonist (Vivo) Negative UT Call (VivoGL)
Dicofol (115-32-2)	Very Weak Agonist (Vitro)
Fenarimol (60168-88-9)	Very Weak Agonist (Vitro)
Spironolactone (52-01-7)	Inactive Agonist (Vitro)
Corticosterone (50-22-6)	Inactive Agonist (Vitro)

Structure	Name (CASRN)	Reported Agonist Activity
Kepone (143-50-0)	Weak Agonist (Vitro) Positive UT Call (VivoGL)
Kaempferol (520-18-3)	Very Weak Agonist (Vitro) Inactive Agonist (Vivo) Negative UT Call (VivoGL)
4-Hydroxybenzoic acid (99-96-7)	Inactive Agonist (Vivo)
Octamethylcyclotetrasiloxane (556-67-2)	Active Agonist (Vivo)
Triclosan (3380-34-5)	Positive UT Call (VivoGL)
Pendimethalin (40487-42-1)	Positive UT Call (VivoGL)
Benzyl salicylate (118-58-1)	Positive UT Call (VivoGL)

Structure	Name (CASRN)	Reported Agonist Activity
β -hexachlorocyclohexane (319-85-7)	Positive UT Call (VivoGL)
Heptylparaben (1085-12-7)	Negative UT Call (VivoGL)
Tetrabromobisphenol A (79-94-7)	Negative UT Call (VivoGL)
Bis(2-ethylhexyl) terephthalate (6422-86-2)	Negative UT Call (VivoGL)
Phenolphthalein (77-09-8)	Negative UT Call (VivoGL)
Oxybenzone (131-57-7)	Negative UT Call (VivoGL)
2,4-Di-tert-butylphenol (96-76-4)	Negative UT Call (VivoGL)

Abbreviations: Vitro = *in vitro* reference chemicals ^{17, 40}, Vivo = *in vivo* reference chemicals ^{2, 17, 40}, VivoGL = *in vivo* guideline-like chemicals ^{2, 17}.