EPA Public Access

Author manuscript

Regul Toxicol Pharmacol. Author manuscript; available in PMC 2018 December 01.

About author manuscripts

Submit a manuscript

Published in final edited form as:

Regul Toxicol Pharmacol. 2017 December ; 91: 39–49. doi:10.1016/j.yrtph.2017.09.022.

On Selecting a Minimal Set of *In Vitro* Assays to Reliably Determine Estrogen Agonist Activity

Richard Judson¹, Keith Houck¹, Eric Watt¹, and Russell S. Thomas¹

¹U.S. Environmental Protection Agency, RTP NC

Abstract

The US EPA is charged with screening chemicals for their ability to be endocrine disruptors through interaction with the estrogen, androgen and thyroid axes. The agency is exploring the use of high-throughput *in vitro* assays to use in the Endocrine Disruptor Screening Program (EDSP), potentially as replacements for lower-throughput *in vitro* and *in vivo* tests. The first replacement is an integrated computational and experimental model for estrogen receptor (ER) activity, to be used as an alternative to the EDSP Tier 1 *in vitro* ER binding and transactivation assays and the *in vivo* uterotrophic bioassay. The ER agonist model uses a set of 16 *in vitro* assays that incorporate multiple technologies and cell lines and probe multiple points in the ER pathway. Here, we demonstrate that subsets of assays with as few as 4 assays can predict the activity of all 1811 chemicals tested with accuracy equivalent to that of the full 16-assay model. The prediction accuracy against reference chemicals is higher than that of the full chemical set, partly because the larger set contains many chemicals that can cause a variety of types of assay interference. There are multiple accurate assay subsets, allowing flexibility in the construction of a multiplexed assay battery. We also discuss the issue of challenging chemicals, i.e. those that can give false positive results in certain assays, and could hence be more problematic when only a few assays are used.

Keywords

endocrine disruption; estrogen receptor; alternative testing methods; in vitro

Introduction

There are tens of thousands of man-made chemicals in the environment, and some fraction of these are known or suspected to be xenoestrogens, i.e. they mimic the activity of natural estrogens by binding the estrogen receptor (ER) and triggering downstream gene activation (Safe 1995). Such chemicals are potentially endocrine active *in vivo* and could mimic certain actions of natural estrogens. If sufficiently high exposure to such chemicals results in exposures at target ERs *in vivo* that overlaps with their potency for the ER, a variety of adverse outcomes at specific life stages may occur, ranging from birth defects to increased rates of cancer progression. We have developed a computational network model (Judson,

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views of policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Magpantay et al. 2015) based on a large set of *in vitro* ER assays (18 in total), that probe a variety of points in the ER pathway, using both cell-free and cell based assays and a range of detection technologies. The agonist model, which uses data from 16 of the *in vitro* assays, was validated against *in vitro* reference chemical activity, as well as reproducible results from guideline-like *in vivo* uterotrophic studies (Browne, Judson et al. 2015, Kleinstreuer, Ceger et al. 2015). Note that the original model used 18 assays, which included two that were antagonist-specific. The current work focuses on agonist mode activity because that is of greater concern from an environmental perspective.) Also note that the uterotrophic assay is itself a screening assay, and positive activity does not necessarily imply *in vivo* adversity. The performance of the ER agonist model was deemed accurate enough that the US EPA has proposed accepting the model results as an alternative to the Endocrine Disruptor Screening Program (EDSP) Tier 1 *in vitro* ER binding and transactivation assays, and the *in vivo* uterotrophic assay (U.S. EPA 2015).

One limitation of the full ER model is the large number of assays used in it, and an obvious question is whether one could achieve equivalent performance with only a subset of the 16 *in vitro* assays. Within this set of 16 *in vitro* assays, there are 3 cell-free binding assays, 6 receptor dimerization assays, 2 transactivation assays measuring reporter RNA levels, 2 measuring reporter protein levels, and a cell proliferation assay. The main reason for exploring the use of many *in vitro* assays, and developing the full ER network model, is that we recognized that all *in vitro* assays yield false positive and false negative results (Judson, Houck et al. 2016). Additionally, these false activities are not random, but are often due to technology-dependent assay interference (Baell and Holloway 2010, Bruns and Watson 2012). A major cause of false positives is activity triggered by cell stress at concentrations approaching where cytotoxicity occurs (Judson, Houck et al. 2016). The full ER network model was developed to help distinguish true activity from false signals. The key output of the model for a chemical was a value referred to as the ER Agonist Area Under the Curve (AUC) score, which is a value from zero to one, normalized with respect to the positive control 17 β -Estradiol. A chemical with a score close to 1 is likely to be a strong agonist and one with a score close to zero is unlikely to have any ER agonist activity. The computational model additionally assigned scores to a variety of *in vitro* assay interference modes, which could be specific to assays or key biological events. An important point about the full model is that it is a so-called *ab initio* model, meaning that it makes predictions about pathway activity based on measured activity in individual components of the pathway and on the structure of the pathway. In particular, the model did not result from “fitting” data to some simple mathematical form, nor did it employ any machine learning or training data.

Here we explore two issues important to the screening and identification of xenoestrogens: 1) identifying predictive combinations of a minimum number of *in vitro* assays that perform almost equivalently to the full 16-assay model; and 2) identifying chemicals that are not true ER agonists, but are active in a subset of assays. Using these chemicals in addition to the standard reference chemicals can allow more robust characterization of assay and model performance. We demonstrate that a variety of subsets of *in vitro* assays perform as well as the full set of 16 assays measuring ER agonist activity. In addition, we find that the *in vitro* reference chemicals typically used in ER validation studies to date are relatively easy to

classify and tend to make the performance of a given *in vitro* assay subset appear better than it will be for a large collection of chemicals.

Methods

The full ER computational network model and its validation are described elsewhere (Browne, Judson et al. 2015, Judson, Magpantay et al. 2015, Kleinstreuer, Ceger et al. 2015). We developed 65,535 “subset models” for all combinations of 1 to 16 *in vitro* assays. The *in vitro* assays are listed in Table 1, and are further described elsewhere (Judson, Magpantay et al. 2015). (Note that the Tox21_ERa_LUC_VM7_Agonist assay has previously been annotated as being run in BG1 cells, but this cell line has recently been shown to be of MCF7 origin and renamed VM7Luc4E2 ER TA (see Corrigendum to (Brennan, Bassal et al. 2016)). The total database consisted of concentration-response data for the 16 ER *in vitro* assays (only focusing on agonist activity) for a total of 1811 chemicals. The concentration response data were fit to a Hill model to yield a hit call (yes or no), an AC50 (concentration at half maximum activity) and a maximum activity (top or T). Here T values >1 were set to 1 (just as in the full model). Assay-chemical pair AUC values were calculated, equal to

$$AUC(chemical, assay_i) = -\log_{10}(AC50_i) \times T(i \in [1, 16]) \quad (1)$$

Chemical-assay AUC values were scaled so that the median value for the non-zero values for each assay were set to 1.0, and then normalized such that all values were divided by the maximum value across all assay-chemical combinations in the subset model. All individual values were then in the range [0,1], just like the full ER model agonist AUC values. The process of fitting a Hill model to each curve was repeated 1000 times using a smoothed nonparametric resampling (smooth bootstrap). Here points from the original concentration-response series for each chemical assay pair are sampled with replacement, with a small amount of noise added to smooth the results. The noise process was assumed to be normal with a standard deviation equal to the baseline median absolute deviation (bmad), which is estimated from the lowest two concentrations for each assay across all chemicals. The full ER model was then run 1000 times to generated distribution of AUC values. More details of this approach, including comparison of other bootstrapping methods, will be provided in a forthcoming manuscript. (Watt et al. in preparation)The chemical-assay AUC matrices (median, upper and lower 95% confidence interval values) are given in Supplemental files S1a, S1b and S1c.

For each *in vitro* assay combination (subset model), we then fit a linear model with zero intercept of the form

$$AUC_{subset}(chemical) = \sum_{i=1}^N c_i \times AUC(chemical, assay_i) \quad (2)$$

The values of the scaling coefficients (c_i) are optimized to minimize the root mean square error (RMSE) between the subset AUC values and the AUC values from the full model, for

chemicals with the full model $AUC > \text{cutoff}$ (see below). The subset AUC values were separately calculated with the median values for the assay AUCs, from the bootstrapping process, and the lower and upper 95% confidence level AUCs, which provides a confidence interval around the subset AUC value. Based on the knowledge that significant *in vitro* assay interference will often be manifested as activity in only one or two assays, we set the subset model value to zero if only one assay is active. The only exception to this is if we are testing models with only a single *in vitro* assay. For each of the 65,535 combinations of *in vitro* assays, we recorded the RMSE and R^2 for the complete set of chemicals with activity in at least one assay, and separately for the reference chemicals (see below). Additionally, we dichotomized the data, calling the chemical “active” if the value of $AUC_{\text{subset}}(\text{chemical})$ was greater than a specified cutoff, and “inactive” otherwise. The value of the cutoff was determined by performing a scan for the 16-assay subset model vs. the full model, and maximizing the balanced accuracy (average of sensitivity and specificity). This yielded a cutoff value of 0.1, which was the same threshold used in the full model. Based on the dichotomized calls, we then calculated sensitivity, specificity and balanced accuracy for each subset model. No separate class of “equivocal” response was used as in Browne et al. (Browne, Judson et al. 2015). Chemicals are classified as “hard to predict” when the difference between the assay AUC (normalized to be in the range of 0 to 1) and full AUC values is large. For certain analyses, we calculate the average difference between the full AUC and the median assay-chemical AUC. Results for the entire set of 65,535 conditions are given in Supplemental File S2.

All computations were carried out using the R programming language, version 3.6.1 (Ihaka and Gentleman 1996). All code and input files are available from the authors at the web address: ftp://newftp.epa.gov/COMPTOX/STAFF/rjudson/publications/ER_assay_select/.

Here, we evaluate subset models against both the full set of 1811 chemicals with data for the full model, and against a set of reference chemicals. These are chemicals with known ER agonist activity, based on expert review. As we will show in the results section, the reference chemicals tend to be well predicted by many individual assays partially because the assays were optimized against at least some of these reference chemicals. For this study, we used a slightly different set of reference chemicals relative to what was used in the full model paper. We eliminated the *in vitro* antagonist reference chemicals, because the current analysis focused only on agonist activity, and we included a set of *in vivo* reference chemicals used in evaluating the full model against the *in vivo* uterotrophic test (Browne, Judson et al. 2015, Kleinstreuer, Ceger et al. 2015). The full set of reference chemicals are given in Table 2 with their *in vitro* and *in vivo* calls. These lists are publicly available on the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) web site (<http://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/refchem/index.html>).

As subset models were tested, we performed three separate evaluations against the reference chemicals. In the first, we calculated the sensitivity, specificity and balanced accuracy of the subset model relative to the specific classification calls of the full model. Then we calculated the same performance metrics of the subset model relative to the *in vitro* and *in vivo* reference chemical classifications from the literature in Table 2. These can be different, because the predictions of the full model were not always in perfect agreement with the

literature reference chemical classifications. In particular, some of the “very weak” *in vitro* positive reference chemicals were classified as inactive by the full model. (Note that several of these are also inactive *in vivo*.) In the original paper, we hypothesized that this was due to the maximum testing concentration (100 μ M) not being high enough to see activity in these very weak xenoestrogens (Judson, Magpantay et al. 2015).

Results

Using all 16 *in vitro* agonist assays, the performance of the subset linear-fit model against the full ER pathway mathematical model, across all 1811 chemicals, is $R^2=0.96$ with an RMSE of 0.071 (Figure 1). The sensitivity, specificity and balanced accuracy (BA) are 0.95, 0.95 and 0.95, respectively. At the low end (AUC close to zero), all of the subset models tend to over-predict the AUC values of the full ER network model because the full model moves the score into specific interference (“pseudo-receptor”) modes, and out of the agonist (or antagonist) mode (Judson, Magpantay et al. 2015). For chemicals with a full model AUC above the cutoff, the confidence intervals around the full and subset models mostly overlap with the diagonal (red points). Confidence intervals are calculated using a bootstrap approach described elsewhere (Watt et al. in preparation). The relatively large confidence intervals support the use of a simple additive linear approach (Eq. 2) to explore multiple variations of subset models in a computationally efficient way.

The performance metrics for all subset models (combinations of assays) with a minimum BA >0.92 are given in Supplemental File S3. The “minimum BA” is the minimum value for the BA for: *in vitro* literature-based reference chemicals (Judson, Magpantay et al. 2015), *in vivo* literature-based reference chemicals (Kleinstreuer, Ceger et al. 2015), all reference chemicals vs. the full model, and all chemicals vs. the full model. The threshold of 0.92 was selected because it is only slightly below the maximum “minimum BA” for all 16 assays (0.94), but allows for an exploration of a diverse set of models with a range of assay subsets. There are 518 models (out of a total of 65,535) with a minimum BA>0.92, with the number of assays ranging from 3 to 12.

Figure 2 shows the BA for all models as a function of the number of *in vitro* assays used. From examining the top panel for BA for all chemicals, one can see that it is possible to achieve 96% BA (all chemicals vs. the full model) for certain combinations of 7 or more *in vitro* assays. This performance is above that for the model with all 16 assays. (This is likely because the model is being forced to fit more assay noise.) By looking at the bottom panel, one can see that the performance for just the *in vitro* reference chemicals is better on average, and in fact shows that there are sets of as few as 4 assays that correctly classify all of the *in vitro* reference chemicals, relative to their classification in the full model. However, as we explore below, subset models that optimize one metric (e.g. classification of the reference chemicals vs. the full model), may have suboptimal (and potentially very poor) performance on other metrics, such as classification of non-reference chemicals, or classification of reference chemicals against the literature results (Browne, Judson et al. 2015, Kleinstreuer, Ceger et al. 2015).

Table 3 gives the performance metrics for the best model for each number of assays from 1 to 16, where “best” means the subset model with the highest minimum-BA, as defined above. (Note that there were some ties for the best value, so only one example is given here. Values for all models are given in Supplemental File S2.) We see subset models reaching the highest value of 0.94 with as few as 4 assays (in this case the human cell free binding assay NVS_NR_hER, the OT protein dimerization assay OT_ER_ERaERb_1440, the ATG TRANS reporter assay ATG_ERa_TRANS, and the proliferation assay, ACEA_T47D_80hr_Positive). Other combinations with more assays achieve the same level of accuracy, but these typically add to this first set of 4 assays, although the selection of which protein complementation assay(s) are used can vary. Optimal performance drops off as more than 12 assays are used. Notice that the weights of certain assays (usually from OT) are negative, indicating that they are being used to pull down the AUC values, and essentially to compensate for over-sensitivity in other assays. Figure 3 graphically shows this increase and then decrease in performance with the number of assays included. By comparing Figures 2 and 3, we see that some subset models can achieve high BA against the full model, but these models may not be as good overall at simultaneously achieving high accuracy against all of the performance metrics.

Table 4 gives the frequency for selection of the assays in the 512 subset models with minimum BA >0.92. The ATG TRANS assays is selected in almost all models (as it is selected in all of the best subset models in Table 3). This assay tends to have a high sensitivity with a lower specificity (Judson, Magpantay et al. 2015), but is likely included in the subset models to maximize the combined sensitivity. The inclusion of other less sensitive but more specific assays then helps to compensate by bringing the model’s overall rate of false positives and false negatives down. The corresponding ATG CIS assay is rarely used, likely because it has similarly high sensitivity and low specificity characteristics, but to a lesser extent. The ACEA cell proliferation assay is the second most often used. As with the ATG assays, it tends to be highly sensitive, with a low specificity. The next three assays in order are the human, mouse and bovine cell-free binding assays. These typically have a lower sensitivity, missing many weak compounds, but are very specific (few false positives). The OT protein complementation assays are all used at intermediate frequencies. These assays are highly correlated, so that subset models with one of these assays will demonstrate similar performance as with any of the others. The two-hour OT chromatin binding assay (OT_ERa_EREGFP_0120) is almost never selected, perhaps because activity has not stabilized at this short time point. The Tox21 beta lactamase transactivation assay is one of the less frequently selected assays, while the Tox21 VM7 luciferase assay is more frequently used due to its increased sensitivity.

The principal take-home message from the preceding analysis is that there are a variety of *in vitro* assay subsets that can be used to accurately classify chemicals for ER agonist activity. However, none of the models gives exactly the same classification as the full model, so it is worth exploring the issue of what types of chemicals yield false positive results in these assays and models. Note that the false negative reference chemicals in the full model (and in the subset models) are almost all due to the limited upper testing concentration (~100 μ M). Here, we focus on “subset false positive chemicals”, which are ones where the subset model indicates that the chemical is positive for estrogenicity, while the full model indicates that it

is negative. As fewer and fewer assays are included in a model, a strong false positive signal in one assay can lead to a positive overall call, one that would be diluted if more (presumably negative) assays were included. Here, we calculate a metric which is the difference between the largest individual assay AUC (Eq. 1) and the full model AUC. Figure 4 shows a heatmap of all chemicals with a difference of 0.35 or greater, as an illustration. The top three chemicals are selective estrogen receptor modulators that are usually functionally classified as ER antagonists. These have high full-model antagonist AUC scores (Judson, Magpantay et al. 2015), and are expected to have significant activity in the upstream assays in the pathway (binding, protein dimerization, chromatin binding), but limited to no activity in the downstream agonist assays (transactivation, proliferation) (Judson, Magpantay et al. 2015), which is what is seen here. In the second block are several progesterone or glucocorticoid-activating chemicals, active mostly in the ACEA cell proliferation assay. This assay, run in T47D cells, is known to be sensitive to these classes of chemicals independent of the estrogen receptor. A number of the remaining chemicals are detergents or solvents which may be activating the *in vitro* assays through a cell-stress or technology interference pathways. Most of these “false positive” chemicals are principally active in the ATG assays, as observed previously (Judson, Magpantay et al. 2015). One hypothesis about these chemicals is that they are actually estrogenic, but only after bioactivation. The ATG assays are run in a sub-strain of HepG2 cells that was selected for increased CYP activity.

Discussion

In this paper, we have described an approach for selecting a minimum battery of ER assays to use in classifying the estrogenic agonist activity of a chemical. We started with a set of 16 assays covering multiple points in the estrogen signaling pathway, which had been combined using a mathematical model that was used to distinguish true agonist or antagonist activity from assay or technology-specific assay interference activity (Judson, Magpantay et al. 2015), and was validated as an alternative to the rodent uterotrophic assay (Browne, Judson et al. 2015). Here, we showed that one can use a simple linear model that combines individual chemical-assay AUC values to achieve comparable performance for agonist activity. The chemical-assay AUC value is a combination of potency and efficacy (Equation 1). Using this simple linear model, we evaluated the accuracy of all combinations of 1 to 16 assays (65,535 in total) against the results of the full model, and well-studied *in vitro* and *in vivo* reference chemicals.

From this analysis, we found that there were many combinations of subsets of assays that performed at levels comparable to the full 16-assay model. Interestingly, the performance against just the *in vitro* reference chemicals was often (but not always) higher than against either the *in vivo* reference chemicals, or against the complete set of ~1800 chemicals evaluated in the full model. This indicates that a battery evaluated (or validated) against only the standard reference chemicals may overstate the true accuracy of the model. This is due to the fact that there are certain chemicals that will be positive in only one assay or technology, either through assay interference (Bael and Holloway 2010, Bruns and Watson 2012), or through true activity that is only captured by that assay (e.g. through metabolic activation which does not occur in most of the current *in vitro* assays). The assay-specific false positive

activity is a key factor determining the domain of applicability of the assay, which is one aspect of standard assay validation practice (ICCVAM 1997, Hartung, Bremer et al. 2004, Judson, Kavlock et al. 2013). In general, the narrower the set of chemicals used to evaluate an assay, the narrower will be the domain of applicability. We analyzed the types of chemicals that show significant differences between the full model AUC and the AUC for individual assays. This analysis showed that such chemicals include antagonists (which are appropriately active in a subset of the agonist assays), chemicals such as detergents, which are likely causing false activity through cell stress in certain cell types (Judson, Houck et al. 2016), and chemicals only active in metabolically competent cells.

Note that some of the false positive (and false negative) activity in the data set used here is not “true”, i.e. some of it is due to experimental variability at the time the assays were run. However, by using data from such a large set of chemicals, such noise should be overwhelmed by the larger signal.

Based on these results, we can conclude that one could use any one of multiple subset models to accurately predict the estrogenic potential of a chemical. There are subsets that include as few as 4 of the original 16 assays that simultaneously have acceptable performance against the full model, and the *in vitro* and *in vivo* reference chemicals. There are other subsets that have good performance against just the *in vitro* reference chemicals, but perform poorly against the full set, and these subsets would not be acceptable substitutes. The acceptable subsets all have assays that probe diverse points in the ER pathway, and use diverse assay reporting technologies and cell types. Note that there are subsets of assays that are more or less sensitive or specific, so that one could tune a battery, for instance to be more sensitive and less specific, and then follow-up actives with other more specific tests. What we have described here are scientific findings that could be used for detecting *in vitro* ER activity, regardless of the regulatory context. To date, the US EPA has not issued an opinion or guidance on using the subset models in a regulatory context.

Because we have tested a large and diverse library of chemicals in this set of 16 assays, we can characterize the types of chemicals that give false positive results, and hence can help characterize the domain of applicability for the assays. One could construct similar batteries using assays that are not included in the 16 used here, but our results suggest that to evaluate performance, one would need to test more chemicals than just the standard *in vitro* reference set. There is not a universal set of assay interference chemicals, but there are likely ones that yield false positive results in many assays using the same cell type or readout. Therefore, by mining available data for assays similar to a new one that is proposed for inclusion in an ER battery, one could develop a set of chemicals to help characterize that assays domain of applicability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem.* 2010; 53(7): 2719–2740. [PubMed: 20131845]
- Brennan JC, Bassal A, He G, Denison MS. Development of a recombinant human ovarian (BG1) cell line containing estrogen receptor alpha and beta for improved detection of estrogenic/antiestrogenic chemicals. *Environ Toxicol Chem.* 2016; 35(1):91–100. [PubMed: 26139245]
- 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):8804–8814. [PubMed: 26066997]
- Bruns RF, Watson IA. Rules for identifying potentially reactive or promiscuous compounds. *J Med Chem.* 2012; 55(22):9763–9772. [PubMed: 23061697]
- Hartung T, Bremer S, Casati S, Coecke S, Corvi R, Fortaner S, Gribaldo L, Halder M, Hoffmann S, Roi AJ, Prieto P, Sabbioni E, Scott L, Worth A, Zuang V. A modular approach to the ECVAM principles on test validity. *Altern Lab Anim.* 2004; 32(5):467–472. [PubMed: 15656771]
- ICCVAM. Validation and Regulatory Acceptance of Toxicological Test Methods: A Report of the ad hoc Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM). Research Triangle Park, NC, USA: NIEHS; 1997.
- Ihaka R, Gentleman R. R: A language for data analysis and graphics. *J Comput and Graphical Statistics.* 1996; 5:299–314.
- Judson R, Houck K, Martin M, Richard AM, Knudsen TB, Shah I, Little S, Wambaugh J, Woodrow Setzer R, Kothya P, Phuong J, Filer D, Smith D, Reif D, Rotroff D, Kleinstreuer N, Sipes N, Xia M, Huang R, Crofton K, Thomas RS. Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. *Toxicol Sci.* 2016; 152(2):323–339. [PubMed: 27208079]
- Judson R, Kavlock R, Martin M, Reif D, Houck K, Knudsen T, Richard A, Tice RR, Whelan M, Xia M, Huang R, Austin C, Daston G, Hartung T, Fowle JR 3rd, Wooge W, Tong W, Dix D. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. *ALTEX.* 2013; 30(1):51–56. [PubMed: 23338806]
- 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):137–154. [PubMed: 26272952]
- Kleinstreuer NC, Ceger PC, Allen DG, Strickland J, Chang X, Hamm JT, Casey WM. A Curated Database of Rodent Uterotrophic Bioactivity. *Environ Health Perspect.* 2015
- Safe SH. Environmental and dietary estrogens and human health: is there a problem? *Environ Health Perspect.* 1995; 103(4):346–351. [PubMed: 7607134]
- U.S. EPA. Use of High Throughput Assays and Computational Tools; Endocrine Disruptor Screening Program; Notice of Availability and Opportunity for Comment. Washington DC, U.S.: Environmental Protection Agency; 2015. URL: <https://www.federalregister.gov/documents/2015/06/19/2015-15182/use-of-high-throughput-assays-and-computational-tools-endocrine-disruptor-screening-program-notice>

Figure 1.

Comparison of the ER agonist AUC scores for the full network model and the subset model with all 16 *in vitro* assays for 1811 chemicals. Confidence intervals calculated using the method described in Watt et al. (in preparation). Red circles indicate that the subset model is within the confidence intervals of the full network model. Sensitivity, specificity and BA for the 16 assay model are 0.95, 0.95 and 0.95 for all chemicals, and 0.97, 0.95 and 0.96 for the *in vitro* reference chemicals. R^2 values are 0.96 for all chemicals and 0.98 for the *in vitro* reference chemicals. The horizontal and vertical lines indicate the cutoff of 0.1.

Figure 2.

Boxplots of balanced accuracy (vs. the full model, using a cutoff of 0.1) for all subset models as a function of the number of *in vitro* assays used, (A) all chemicals and (B) *in vitro* reference chemicals. The bar in the middle of the box is the median value, the box indicates the middle two quantiles, the whiskers the 95% level, and points outliers beyond 95% of the distribution. The horizontal line is set at 0.96, the maximum BA seen for any model across all chemicals. Statistics for all individual models are given in Supplemental File S2.

Figure 3.

Performance for the best models for each number of *in vitro* assays from 1 to 16. The black symbols indicate data for the best all-chemical model, corresponding to Table 3. The red symbols are shifted to the right slightly for ease of visualization.

Figure 4.

Heatmap of the maximum assay subset AUC values for chemicals with absolute difference between the maximum assay subset AUC and the full model AUC of >0.35 . The color scale ranges from white (0.0) to dark red (1.0). The AUC could be for any subset with fewer than 16 assays.

Summary of the *in vitro* ER agonist assays used in the analysis. Further details are provided in the primary reference for the ER network model.

Table 1

Assay ID	Assay Name	Source ^[1]	Gene Target	Normalized Data Type	Assay Design Type	Biological Process Subtype	Detection Technology	Detection Technology Subtype	Measurement Time Point (hr)	Organism	Tissue	Cell Format	Cell Line Name	Assay Footprint
A1	NVS_NR_bER	NVS	ESR1	percent activity	radioligand binding	receptor binding	Lysate-based radiodetection	Scintillation counting	18	bovine	uterus	tissuebased cell-free	NA	microplate: 96-well plate
A2	NVS_NR_hER	NVS	ESR1	percent activity	radioligand binding	receptor binding	Lysate-based radiodetection	Scintillation counting	18	human	NA	cell-free	NA	microplate: 96-well plate
A3	NVS_NR_mERa	NVS	Esr1	percent activity	radioligand binding	receptor binding	Filter-based radiodetection	Scintillation counting	18	mouse	NA	cell-free	NA	microplate: 96-well plate
A4	OT_ER_ERaERa_0480	OT	ESR1	percent activity	protein fragment complementation assay	protein stabilization	Protein-fragment complementation	Fluorescence intensity	8	human	kidney	cell line	HEK293T	microplate: 384-well plate
A5	OT_ER_ERaERa_1440	OT	ESR1	percent activity	protein fragment complementation assay	protein stabilization	Protein-fragment complementation	Fluorescence intensity	24	human	kidney	cell line	HEK293T	microplate: 384-well plate
A6	OT_ER_ERaERb_0480	OT	ESR1	percent activity	protein fragment complementation assay	protein stabilization	Protein-fragment complementation	Fluorescence intensity	8	human	kidney	cell line	HEK293T	microplate: 384-well plate
A7	OT_ER_ERaERb_1440	OT	ESR1 ESR2	percent activity	protein fragment complementation assay	protein stabilization	Protein-fragment complementation	Fluorescence intensity	24	human	kidney	cell line	HEK293T	microplate: 384-well plate
A8	OT_ER_ERbERb_0480	OT	ESR2	percent activity	protein fragment complementation assay	protein stabilization	Protein-fragment complementation	Fluorescence intensity	8	human	kidney	cell line	HEK293T	microplate: 384-well plate
A9	OT_ER_ERbERb_1440	OT	ESR2	percent activity	protein fragment complementation assay	protein stabilization	Protein-fragment complementation	Fluorescence intensity	24	human	kidney	cell line	HEK293T	microplate: 384-well plate
A10	OT_ERa_EREGFP_0120	OT	ESR1	percent activity	fluorescent protein induction	regulation of gene expression	Microscopy	Optical microscopy: Fluorescence microscopy	2	human	cervix	cell line	HeLa	microplate: 384-well plate
A11	OT_ERa_EREGFP_0480	OT	ESR1	percent activity	fluorescent protein induction	regulation of gene expression	Microscopy	Optical microscopy: Fluorescence microscopy	8	human	cervix	cell line	HeLa	microplate: 384-well plate
A12	ATG_ERa_TRANS_up	ATG	ESR1	log2 fold induction	mRNA induction	regulation of transcription factor activity	RT-PCR and Capillary electrophoresis	Fluorescence intensity	24	human	liver	cell line	HepG2	microplate: 24-well plate
A13	ATG_ERE_CIS_up	ATG	ESR1	log2 fold induction	mRNA induction	regulation of transcription factor activity	RT-PCR and Capillary electrophoresis	Fluorescence intensity	24	human	liver	cell line	HepG2	microplate: 24-well plate
A14	Tox21_ERa_BLA_Agonist_rat o	Tox21	ESR1	percent activity	beta lactamase induction	regulation of gene expression	GAL4 b-lactamase reporter gene	Fluorescence intensity	18	human	kidney	cell line	HEK293T	microplate: 1536-well plate
A15	Tox21_ERa_LUC_VMT_Agonist	Tox21	ESR1	percent activity	luciferase induction	regulation of gene expression	Luciferase-coupled ATP quantitation	Bioluminescence	22-24	human	ovary	cell line	VM7	microplate: 1536-well plate
A16	ACEA_T47D_80hr_Positive	ACEA	ESR1	percent activity	real-time cell-growth kinetics	cell proliferation	RT-CES	Electrical Sensor: Impedance	80	human	breast	cell line	T47D	microplate: 96-well plate

^[1]NVS=Novascreen; OT=Odyssey Thera; ATG=Attagene; Tox21=assays run by the National Institutes of Health’s National Center for Advancing Translational Sciences (NCATS) as part of the Federal Tox21 program.

Table 2

Reference chemicals from NICEATM

CASRN	Name	<i>In Vitro</i> Agonist Activity Class	<i>In Vivo</i> Agonist Activity Class
57-91-0	17alpha-Estradiol	Moderate	Active
131-56-6	2,4-Dihydroxybenzophenone		Active
140-66-9	4-(1,1,3,3-Tetramethylbutyl)phenol	Moderate	Active
80-46-6	4-(1,1-Dimethylpropyl)phenol		Active
104-43-8	4-Dodecylphenol		Active
99-96-7	4-Hydroxybenzoic acid		Inactive
521-18-6	5alpha-Dihydrotestosterone	Weak	Active
61-82-5	Amitrole		Inactive
520-36-5	Apigenin	Very Weak	
5153-25-3	Benzoic acid, 4-hydroxy-, 2-ethylhexyl ester		Active
131-55-5	Benzophenone-2		Active
103-23-1	Bis(2-ethylhexyl)hexanedioate		Inactive
80-05-7	Bisphenol A	Weak	Active
1478-61-1	Bisphenol AF		Active
77-40-7	Bisphenol B	Weak	Active
80-09-1	Bisphenol S		Active
85-68-7	Butylbenzyl phthalate	Very Weak	
94-26-8	Butylparaben		Active
480-40-0	Chrysin	Very Weak	
50-22-6	Corticosterone	Inactive	
486-66-8	Daidzein	Weak	
115-32-2	Dicofol	Very Weak	
84-61-7	Dicyclohexyl phthalate		Inactive
84-66-2	Diethyl phthalate		Inactive
117-81-7	Diethylhexyl phthalate	Very Weak	Inactive
56-53-1	Diethylstilbestrol (DES)	Strong	Active
84-75-3	Dihexyl phthalate		Inactive
84-74-2	Di-n-butyl phthalate	Very Weak	Inactive
474-86-2	Equilin		Active
50-28-2	Estradiol	Strong	Active
50-27-1	Estriol		Active
53-16-7	Estrone	Moderate	Active
57-63-6	Ethinyl Estradiol	Strong	Active
120-47-8	Ethylparaben	Very Weak	
60168-88-9	Fenarimol	Very Weak	
51630-58-1	Fenvalerate		Inactive
446-72-0	Genistein	Weak	Active

CASRN	Name	<i>In Vitro</i> Agonist Activity Class	<i>In Vivo</i> Agonist Activity Class
52-86-8	Haloperidol	Inactive	
520-18-3	Kaempferol	Very Weak	Inactive
143-50-0	Kepone	Weak	
65277-42-1	Ketoconazole	Inactive	
330-55-2	Linuron	Inactive	
84-16-2	meso-Hexestrol	Strong	
72-33-3	Mestranol		Active
72-43-5	Methoxychlor	Very Weak	Active
58-18-4	Methyltestosterone	Very Weak	Active
104-40-5	Nonylphenol	Very Weak	Active
68-22-4	Norethindrone		Active
789-02-6	o,p'-DDT	Weak	Active
556-67-2	Octamethylcyclotetrasiloxane		Active
72-55-9	p,p'-DDE	Very Weak	
599-64-4	p-Cumylphenol	Weak	Active
87-86-5	Pentachlorophenol		Inactive
57-30-7	Phenobarbital Sodium	Inactive	
32809-16-8	Procymidone	Inactive	
98-54-4	p-tert-Butylphenol		Active
50-55-5	Reserpine	Inactive	
52-01-7	Spironolactone	Inactive	
17924-92-4	Zearalenone		Active

Table 3

Best models for each number of assays. Columns (1, 2, 3... 16) indicate the number of assays. BA (balanced accuracy), sensitivity and specificity values are measured relative to the classification of all chemicals in the full model.

Assays	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Sensitivity (all chemicals)	0.94	0.91	0.95	0.98	0.95	0.98	0.95	0.96	0.97	0.97	0.96	0.94	0.91	0.94	0.91	0.95
Specificity (all chemicals)	0.80	0.89	0.90	0.92	0.92	0.92	0.94	0.93	0.94	0.94	0.94	0.91	0.94	0.95	0.94	0.95
BA (all chemicals)	0.87	0.90	0.93	0.95	0.94	0.95	0.94	0.95	0.96	0.96	0.95	0.93	0.92	0.94	0.92	0.95
BA (in vitro reference chemicals)	0.87	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.93	0.95	0.88	0.93	0.93
BA (in vivo reference chemicals)	0.86	0.91	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.89	0.89	0.88	0.88
BA (minimum)	0.86	0.90	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.93	0.89	0.88	0.88	0.88
Assay Selection																
NVS_NR_bER	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
NVS_NR_hER	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
NVS_NR_mERa	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1	1
OT_ER_ERaERa_0480	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
OT_ER_ERaERa_1440	0	0	0	0	1	0	1	1	1	0	0	1	1	1	1	1
OT_ER_ERaERb_0480	0	0	0	0	0	1	1	0	1	1	1	1	1	1	1	1
OT_ER_ERaERb_1440	0	0	1	1	0	1	0	0	0	1	1	1	1	1	1	1
OT_ER_ERbERb_0480	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1	1
OT_ER_ERbERb_1440	0	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1
OT_ERa_EREGFP_0120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
OT_ERa_EREGFP_0480	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1	1
ATG_ERa_TRANS_up	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ATG_ERE_CIS_up	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
TOX21_ERa_BLA_Agonist_ratio	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
TOX21_ERa_LUC_BG1_Agonist	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
ACEA_T47D_80hr_Positive	0	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1
Assay Weights																
NVS_NR_bER	0	0	0	0	0	0	0	0.23	0.22	0.24	0.22	0.28	0.22	0.24	0.18	0.19
NVS_NR_hER	0	0	0.35	0.26	0.18	0.2	0.23	0.04	0.04	0.05	0.06	0	0.01	-0.05	0.02	-0.02

Assays	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
NVS_NR_mERa	0	0	0	0	0.08	0.09	0	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.01	0.01
OT_ER_ERaERa_0480	0	0	0	0	0	0	0.11	-0.02	0.01	0.06	0.06	0.09	0.06	0.04	0.03	0.02
OT_ER_ERaERa_1440	0	0	0	0	0.16	0	0.14	0.07	0.06	0	0	0.07	0.05	0.05	0.07	0.07
OT_ER_ERaERb_0480	0	0	0	0	0	0.03	-0.02	0	-0.03	0.02	0.04	-0.1	-0.09	-0.07	-0.1	-0.09
OT_ER_ERaERb_1440	0	0	0.15	0.13	0	0.08	0	0	0	0.02	-0.03	0.06	0.1	0.1	0.02	0.03
OT_ER_ERbERb_0480	0	0	0	0	0	0	-0.05	0	0	-0.08	-0.11	-0.02	0.05	0.05	0.02	0.03
OT_ER_ERbERb_1440	0	0.37	0	0	0	0	0	0	0	0	0.06	-0.08	-0.1	-0.13	-0.06	-0.09
OT_ERa_EREGFP_0120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.07	0.08
OT_ERa_EREGFP_0480	0	0	0	0	0	0	0	0.16	0.17	0.17	0.18	0	0	0	0.11	0.08
ATG_ERa_TRANS_up	0.67	0.34	0.29	0.23	0.25	0.23	0.23	0.24	0.25	0.24	0.23	0.11	0.05	0.03	0.06	0.05
ATG_ERE_CIS_up	0	0	0	0	0	0	0	0	0	0	0	0	0.17	0.12	0.14	0.1
TOX21_ERa_BLA_Agonist_ratio	0	0	0	0	0	0	0	0	0	0	0	0	0.23	0.25	0.19	0.21
TOX21_ERa_LUC_BG1_Agonist	0	0	0	0	0	0	0	0	0	0	0	0.34	0.19	0.18	0.22	0.2
ACEA_T47D_80hr_Positive	0	0	0	0.15	0.14	0.16	0.15	0.15	0.15	0.15	0.15	0.13	0	0.12	0	0.09

Table 4

Frequency of use of assays in models with minimum BA>0.92

Assay	Frequency
ATG_ERa_TRANS_up	0.98
ACEA_T47D_80hr_Positive	0.82
NVS_NR_hER	0.77
NVS_NR_mERa	0.57
NVS_NR_bER	0.53
OT_ER_ERaERb_0480	0.51
TOX21_ERa_LUC_VM7_Agonist	0.47
OT_ER_ERaERb_1440	0.46
OT_ER_ERbERb_1440	0.42
OT_ER_ERbERb_0480	0.42
OT_ERa_EREGFP_0480	0.37
OT_ER_ERaERa_1440	0.32
OT_ER_ERaERa_0480	0.31
TOX21_ERa_BLA_Agonist_ratio	0.23
ATG_ERE_CIS_up	0.13
OT_ERa_EREGFP_0120	0.01