



Evaluation of OASIS QSAR Models Using ToxCast
in Vitro Estrogen and Androgen Receptor Binding Data
and Application in an Integrated Endocrine Screening
Approach

Barun Bhatarai, Daniel M. Wilson, Paul S. Price, Sue Marty,
Amanda K. Parks, and Edward Carney

<http://dx.doi.org/10.1289/EHP184>

Received: 29 July 2015
Revised: 31 December 2015
Accepted: 22 April 2016
Published: 6 May 2016

Note to readers with disabilities: *EHP* will provide a [508-conformant](#) version of this article upon final publication. If you require a 508-conformant version before then, please contact ehp508@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

Evaluation of OASIS QSAR Models Using ToxCast *in Vitro* Estrogen and Androgen Receptor Binding Data and Application in an Integrated Endocrine Screening Approach

Barun Bhatarai, Daniel M. Wilson, Paul S. Price, Sue Marty, Amanda K. Parks, and Edward Carney

Toxicology Environmental Research and Consulting, The Dow Chemical Company, Midland, Michigan, USA

Address correspondence: Barun Bhatarai, The Dow Chemical Company, Midland, MI 48674 USA. Telephone: 989-638-6862. Fax: 989-638-9305. E-mail: bbhatarai@dow.com

Running title: Evaluation of QSAR models for use in an Integrated Endocrine Screening

Acknowledgments: The authors gratefully acknowledge Dr. J. Craig Rowlands for critical review of the manuscript and Tyler Auernhammer for data collection. We also like to thank US EPA for recognizing part of this work with an ‘Exemplary Poster Award for Non-Federal Employees’ at the ToxCast Data Summit in Research Triangle Park, NC, in 2014. All authors were employed by The Dow Chemical Company, Midland, MI, USA during the time of the conduct of this research.

Competing financial interests: The authors declare they have no actual or potential competing financial interests. BB performed the research and wrote the manuscript. All authors reviewed and commented on the content. Paul S. Price is now employed by the U.S. EPA but was employed by The Dow Chemical Company during most of this research.

Abstract

Background: Integrative testing strategies (ITS) for potential endocrine activity can use tiered *in silico* and *in vitro* models. Each component of an ITS should be thoroughly assessed.

Objectives: We used the data from three *in vitro* ToxCast binding assays to assess OASIS, a quantitative structure-activity relationship (QSAR) platform covering both estrogen (ER) and androgen receptor (AR) binding. For stronger binders (described here as $AC_{50} < 1 \mu\text{M}$), we also examined the relationship of QSAR predictions of ER or AR binding to the results from 18 ER and 10 AR transactivation assays, 72 ER-binding reference compounds as well as the *in vivo* uterotrophic assay.

Methods: NovaScreen binding assay data for ER (human, bovine and mouse) and AR (human, chimp and rat) were used to assess the sensitivity, specificity, concordance and applicability domain of two OASIS QSAR models. The binding strength relative to the QSAR-predicted binding strength was examined for the ER data. The relationship of QSAR predictions of binding to transactivation and pathway based assays, as well as *in vivo* uterotrophic responses was examined.

Results: The QSAR models had both high sensitivity (>75%) and specificity (>86%) for ER as well as both high sensitivity (92-100%) and specificity (70-81%) for AR. For compounds within the domains of the ER and AR QSAR models that bound with $AC_{50} < 1 \mu\text{M}$, the QSAR models accurately predicted the binding for the parent compounds. These were active in all transactivation assays where metabolism was incorporated and, except for those compounds known to require metabolism to manifest activity, all assay platforms where metabolism wasn't incorporated. Compounds in-domain and predicted to bind by the ER QSAR model that were positive in ToxCast ER binding at $AC_{50} < 1 \mu\text{M}$ were active in the uterotrophic assay.

Conclusions: We used the extensive ToxCast HTS binding data set to show that OASIS ER and AR QSAR models had high sensitivity and specificity when compounds were in-domain of the models. Based on this research, we recommend a tiered screening approach wherein (1) QSAR is used to identify compounds in-domain of the ER or AR binding models and predicted to bind; (2) those compounds are screened *in vitro* to assess binding potency; and (3) the stronger binders ($AC_{50} < 1 \mu\text{M}$) are screened *in vivo*. This scheme prioritizes compounds for integrative testing and risk assessment. Importantly, compounds not in-domain, predicted either not to bind or to bind weakly, not active in *in vitro*, that require metabolism to manifest activity or for which *in vivo* AR testing is in order, need to be assessed differently.

Introduction

The ability to quickly profile and prioritize large numbers of compounds for potential hazards, including endocrine receptor binding activity, improved with the advent of predictive *in vitro* high-throughput screening (HTS) methodologies such as ToxCast (Dix et al. 2007; Kavlock et al. 2012). ToxCast uses a battery of HTS assays to develop activity signatures across a range of *in vitro* endpoints and chemistries. ToxCast phase II included approximately 1800 compounds screened in a subset of assays focusing on potential endocrine activity, which included both biochemical and cell-based measures. Similarly, development of *in silico* predictive QSAR models for toxicity estimations is continuously advancing. Independent QSAR models are available for some of the endpoints in ToxCast, such as estrogen receptor (ER) and androgen receptor (AR) binding, which were originally developed using smaller subsets of *in vitro* receptor binding data generated under other platforms. In this article, we use the extensive ER

and AR assay data published for ToxCast phase II to assess the performance and thus further delineate the validity of the 3D-QSAR model predictions of the ER and AR binding models implemented in OASIS (Mekenyan et al. 2000). The predictions in terms of receptor binding potency are addressed for the ER model. The likelihood of using a tiered approach to flag in-domain compounds predicted as active by QSAR and demonstrated as stronger binders in ToxCast (AC_{50} or concentration at which activity is 50% of its maximum $< 1 \mu\text{M}$) to identify compounds that would also be active in respective transactivation and the uterotrophic assay is addressed.

Assessment of potential endocrine activity of compounds is an area of intense focus world-wide. In the USA, the Endocrine Disruptor Screening Program (EDSP) (EDSP21) has developed a tiered assessment and prioritization scheme to screen compounds in commerce with the potential for consumer exposure. EDSP screening ranges from short-term *in vitro* assays to multigenerational studies in which eleven assays (five *in vitro* and six *in vivo*) are used as a first tier to determine whether compounds interact with three endocrine hormonal pathways – ER, AR and thyroid (EDSP21). Compounds that bind to ER or AR might influence endocrine signaling by either blocking the binding of the endogenous hormones, by activating receptor signaling or both (Katzenellenbogen 1995; Katzenellenbogen et al. 2003). These compounds may also mimic the action of hormones due to their structural similarity and may initiate similar downstream sequelae or alter the concentrations of hormones affecting their synthesis, transport, metabolism and excretion (Katzenellenbogen 1995; Katzenellenbogen et al. 2003). The US-EPA ToxCast program has made *in vitro* HTS data publically available for a broad range of cellular and biochemical targets that cover major protein super families, key signaling pathways and

phenotypic endpoints. In addition, known nuclear receptor (NR) targets, including steroid hormone receptors such as ER and AR, are included (Kavlock et al. 2012).

The ER and AR binding assays studied here were implemented by NovaScreen (NVS) and covered cloned receptors isolated from three different mammalian species. Other assays that include transactivation- and pathway-based assays were from Odyssey Thera (OT), Attagene (ATG), ACEA and Tox21 platforms. Uterotrophic assay data were obtained from a curated dataset (ER platform) as described recently (Browne et al. 2015). To date, the half-maximal activity concentration (AC_{50}) value has been the most commonly used *in vitro* parameter under ToxCast and was used herein (Dix et al. 2007; Kavlock et al. 2012). There are several available *in silico* models for predicting ER and AR activity, which are summarized elsewhere (Piparo and Worth 2010). Recent manuscripts revealed various approaches such as structure-based (Steinmetz et al. 2015), docking-based (Kolsek et al. 2014) or mathematical models using *in vitro* data (Browne et al. 2015; Judson et al. 2015; McRobb et al. 2014; Zhang et al. 2013) to access ER binding. Here, we evaluated both the qualitative and quantitative predictivity of OASIS QSAR model for ER and AR binding and transactivation using the ToxCast HTS screening data as the challenge dataset. The predictions in OASIS are based on the combination of a toxico-dynamic and toxico-kinetic model in a single platform, where a mechanistic QSAR model for ER and AR binding affinity is combined with metabolism models (referred in the QSAR tool as Tissue MEtabolism Simulator (TIMES)) to address binding of either the parent compound or its predicted metabolites (Mekenyan et al. 2000). The Common Reactivity Pattern (COREPA) approach (Bradbury et al. 2000; Mekenyan et al. 2000) is implemented in the software, which helps to identify stereo-electronic characteristics associated with a chemical's biological activity by incorporating dynamic conformational flexibility. For the ER model

(trained on human and trout *in vitro* assay data), relative binding affinity (RBA) is predicted relative to 17 β -estradiol (100% binding). Nucleophilicity, interatomic distance between electronegative heteroatoms and electron donor capability of heteroatoms are all important model variables. Similarly, the AR binding affinity model is based on a set of stereo-electronic parameters that provide a maximal measure of pair-wise similarity among the conformers of the most active steroidal and non-steroidal ligands. The standard QSAR model assesses the binding affinity of the parent compounds only, unless the metabolism is switched on or compounds are evaluated with the metabolism simulator first before running the models (Shelby et al. 1996). Principle metabolic transformations include oxidative reactions like aromatic ring hydroxylation and O-dealkylation that are generated by hepatic cytochrome P450 (CYP) enzymes. In this article, prediction results obtained from the use of ER and AR receptor binding 3D-QSAR models from OASIS-TIMES relative to the ToxCast phase II compounds will be presented. The models were run herein without incorporation of the TIMES QSAR metabolism simulation because the ToxCast binding assays did not incorporate metabolism. Further, for the subset of in-domain compounds predicted and shown to be active across all ToxCast binding platforms with $AC_{50} < 1 \mu M$, the outcome of results in respective transactivation assays and the uterotrophic assay will be presented. The reliability of the QSAR estimate compared to the ToxCast assay result as well as the mechanistic explanation of the QSAR estimation for some compounds will be discussed. The utility of ToxCast data for the refinement of the existing QSAR models will also be suggested.

Methods

Compounds and Assays. All compounds with data on *in vitro* ER and AR assays were obtained from the ToxCast phase II public release on 12/05/2013. (Dix et al. 2007). These assays span

across different species, targets, genes, etc. (Supplemental Material, Table 1). Compounds without defined structures such as oils or mixtures like Milbemectin (containing Milbemcin A4 and Milbemycin A3) were excluded during the evaluation. For compounds with more than one component, salts or acids were removed and only the unique parent compound was predicted by the model. As the bioactivity (AC_{50}) data of the parent and salts were different, for comparison purposes, the original compound name and structure was kept the same as provided by ToxCast. The total chemical lists, CAS numbers, SMILES codes, corresponding ToxCast assay values, potency bins, and calculated RBA values are given as Supplemental Material, Excel Table 2a. In addition, uterotrophic response were obtained from a recent publication (Browne et al. 2015) for 42 compounds that were a subset of the ToxCast dataset. 72 reference chemicals used in developing an integrated model for validating ToxCast ER assays (Judson et al. 2015) were also compared.

QSAR modeling. OASIS (v2.27.13) predictions for receptor mediated endpoints of ER and AR were calculated using estrogen binding affinity (v.03) and androgen binding affinity (v.03) 3D-QSAR models (Mekenyan et al. 2000). The ER model was built with 853 compounds in the training set containing 650 human ER and 153 trout ER relative binding affinity (RBA) data-points (Katzenellenbogen et al. 2003; Serafimova et al. 2007). Similarly, the AR model was built with 202 compounds in the training set with observed RBA based on recombinant rat protein expressed in *Escherichia coli* whose ligand binding domain is considered to be similar to human AR (Fang et al. 2003; Kelce et al. 1994; Mekenyan et al. 1997; Waller et al. 1996). Both ER and AR models are based on cell-free competitive radio-labeled receptor binding *in vitro* assays. The training set compounds were also ranked for RBA for ER and AR. These models were applied in a batch mode to the compounds based on SMILES information published by the US-EPA. Each

of the modeled endpoints was run individually and the results exported as a tab-delimited text file. The AM1 Hamiltonian method for MOPAC calculation and ‘Accurate’ conformer generation was selected. The models were run using only OASIS without simulation of metabolism by TIMES. The applicability domain of the models was studied based on the default values selected for total domain estimation as defined in the program. The total domain is a combination of structural-, mechanistic- and parametric-based domains; compounds not satisfying the criteria in any of the sub-domain makes them out-of-domain in total.

Performance of QSAR models. The performances of the ER and AR QSAR models for ToxCast compounds for individual ER and AR assays related to the mammalian nuclear receptor (NR) targets were studied based on specificity, sensitivity and concordance. In addition, the potency of binding predictions for *in vitro* compounds was calculated based on the RBA compared to the positive control (estradiol for ER), converted into percentile and compared with the respective *in silico* predictions. No potency-based prediction for AR binding was performed as we were unable to obtain AC₅₀ values for the positive control (i.e., R1881). The *in silico* and *in vitro* predictions were assigned into four different bins for ER: RBA greater than 10% of positive control (denoted as High), from 0.1 to 10% (Medium), from 0.001 to 0.1% (Low) and from 0 to 0.001% (Very Low). If the OASIS prediction was uncertain and assigned into two bins for a chemical, the most conservative prediction bin (i.e., highest predicted activity) was chosen. The impact on the probability of being positive a priori in the assay was also calculated using Bayesian statistics. These calculations were done for the in-domain compounds as well as the total compounds.

Heatmaps for ER and AR assays. Heat maps for ER- and AR-related assays were generated using TIBCO Spotfire (SpotFire) to index the representative performance of individual assays for the

subset of compounds that showed consistent agreement of ER or AR binding at $AC_{50} < 1 \mu M$ (stronger binders). Colors were coded to indicate inactive and actives with different ranges of *in vitro* potency.

Results

ER and AR QSAR model performance for the binding assays. The performance of the ER and AR QSAR models for predicting the binding activity of in-domain compounds (by the respective QSAR model) versus the ToxCast *in vitro* binding test results are summarized in Table 1. The top half of each table shows the results for ER compounds whereas the lower half shows the results for the AR compounds. Similar tables for ‘all’ compounds are given in Supplemental Material, Table 2b.

For the ER QSAR model, 458 (24.8%) of 1845 compounds were in-domain and 1365 (74%) were out-of-domain. The model assigned ‘No domain’ for 17 (0.9%) and no information was provided for 5 (0.3%) compounds. For in-domain compounds, the ER QSAR model had both high sensitivity (>75%) and specificity (>86%). ER QSAR predictions had low sensitivity (< 56%) but high specificity (>95%) when the model was applied without distinction of domain boundaries. For the 458 in-domain compounds, 87 (19.0%) were predicted to be active and 371 (81.0%) were predicted to be inactive. For compounds in-domain, the overall concordance decreased by approximately 5% and sensitivity increased by 36-38% compared to predictions for the total compound dataset.

For the AR QSAR model, 213 (12.1%) of 1758 compounds were in-domain and 1516 (86.2%) were out-of-domain. The model assigned ‘No domain’ for 17 (0.9%) and no information was provided for 12 (0.7%) compounds. For in-domain compounds, the AR QSAR model had both

high sensitivity (92-100%) and specificity (70-81%). AR QSAR model predictions had low sensitivity (<41%) but high specificity (84-89%) when the model was applied without distinction of domain boundaries. For the 213 in-domain compounds, 69 (32.4%) were predicted to be active and 144 (67.6%) were predicted to be inactive. For compounds in-domain, the overall concordance decreased by approximately 10% and sensitivity increased by 53-64% compared to predictions for the total compound dataset.

Consideration of stronger ER and AR binders. For both the ER and AR QSAR models, when the HT results for all three mammalian nuclear receptor binding assays were restricted to those showing consistent agreement of binding at $AC_{50} < 1 \mu\text{M}$, the QSAR models accurately predicted binding for the parents or known hydroxylated metabolites 100% of the time. There were 20 compounds for ER (Table 2) and 11 for AR (Table 3) for which the HT results for all three respective binding assays showed consistent agreement of binding at $AC_{50} < 1 \mu\text{M}$. For the ER QSAR model, three compounds (Clomiphene, Tamoxifen and Tamoxifen citrate; Figure 1) were out of the domain of the model but their known mammalian 4-hydroxyphenyl metabolites satisfied the structural domain boundary requirements. It seems apparent that the domain of ER model is constrained to phenols, which is consistent with the fact that the 4-hydroxy-Tamoxifen was in-domain and approximately ten-fold more potent a binder than either Tamoxifen or its citrate salt, which were considered out-of-domain by the model. Thus for the ER model, considering not only the parent compounds but their hydroxylated metabolites was necessary to obtain the 100% prediction. Both parent and hydroxylated metabolites bound strongly ($AC_{50} < 1 \mu\text{M}$) in the ToxCast assays. Two other compounds (Raloxifene hydrochloride and Phenolphthalein) were also strong binders *in vitro* but were predicted to be out-of-domain by the ER QSAR model. Thus, such data could be used to improve the QSAR model. For the AR

QSAR model, all 11 compounds were predicted to be active and in-domain 100% of the time except for Mifepristone which was not predicted and was out-of-domain.

Heatmaps for ER and AR assays. Heatmaps were generated for the subset of compounds that showed consistent agreement of ER or AR binding at ToxCast $AC_{50} < 1 \mu\text{M}$ and where the QSAR models predicted in-domain binding for the parents or known hydroxylated metabolites 100% of the time. The heatmaps (Figure 2) show the distribution of ToxCast *in vitro* activity for each assay and are color-coded by potency. This subset of compounds was active in all ER and AR transactivation assays where metabolism was incorporated with addition of an exogenous S9 fraction (OT platform). For those compounds in this subset known to require metabolism to manifest activity, the transactivation response appears to be less promiscuous than binding because of the mixed nature of the response in these assays, which may reflect some degree of constitutive metabolism depending on the cell type.

For ER, the heatmap (Figure 2) shows the 18 assays selected by the US EPA (Judson et al. 2015) to derive an integrated model for ER pathway perturbation and were also included as component assays of the EDSP21 Dashboard (EDSP21). Figure 2 shows that the ER compounds were inactive in the respective antagonist assays, Era_BLA_Antagonist and Era_LUC_BG1_Antagonist, except for 4-hydroxytamoxifen and raloxifene that are known selective ER modulators (SERMs). Four of the compounds known to undergo further metabolism in relation to ER activity (Tamoxifen and its citrate salt, Clomiphene citrate and Raloxifene Hydrochloride) were inactive in half of the ER assays that didn't include metabolism but active in those that did. 4-OH Tamoxifen was inactive in the assays that didn't include metabolism and active in those that did; this was an interesting observation in view of its known ER binding.

Similarly, for AR, all selected compounds were active in all assays except for two, Tox21_AR_LUC_MDAKB2_Antagonist and Tox21_AR_BLA_Antagonist_ratio, as expected as the compounds were agonists. The ATG_AR_TRANS assay exhibited differential activation to known Progesterone receptor agonists (such as Spironolactone, Mifepristone, Cyproterone) but not all (Norethindrone and Norgestrel) (Figure 3). Our findings highlight the ability to use QSAR models to predict those compounds that were uniformly active in all respective ToxCast assay platforms with the caveat of needing to account for some metabolic activation.

ER QSAR model performance for the uterotrophic assay: We further examined the ability of the ER QSAR model to predict the activity of *in vivo* responses in a curated uterotrophic dataset (Browne et al. 2015). The sensitivity of the QSAR predictions averaged over 80% for those compounds in-domain of the model (Supplemental Material, Table 3a). When the compounds in this group were restricted to the subset of stronger binders ($AC_{50} < 1 \mu M$), the results showed that a QSAR prediction of receptor binding for in-domain compounds identified active compounds of other *in vitro* ER assays and those in the *in vivo* uterotrophic assay. Of note, all the compounds in this restricted group were identified as actives used in the training set of the ER QSAR model (Supplemental Material, Table 3b).

Application of ER QSAR to 72 reference compounds: The selected ToxCast ER-related 18 assays and 72 reference compounds were studied in a recent publication (Judson et al. 2015). These 72 compounds were a subset of ToxCast compounds and were chosen to validate the ER assays. We assessed these compounds and assays using the OASIS QSAR model in the current manuscript. There were 27 reference compounds with $AC_{50} < 1 \mu M$ out of which all except Phenolphthalein were ‘strong binders’ in the ER ToxCast assay list. When the ‘strong binders’ (i.e., $AC_{50} < 1 \mu M$) observed in all three ER binding ToxCast assays were compared with the median AC_{50} values of

the 72 reference compounds (Judson et al. 2015), two strong binders in the ER ToxCast assays list (Daidzein and 2,2',4,4'-Tetrahydroxybenzophenone) had median $AC_{50} > 1 \mu\text{M}$, whereas 10 out of 70 reference compounds with $AC_{50} < 1 \mu\text{M}$ were not 'strong binders' in the ER ToxCast assay list. Compared to the 42 uterotrophic compounds (Browne et al. 2015), there were 25 compounds evaluated in the reference compounds list.

In vitro binding potency estimation. The predicted potency of the *in vitro* ER binding from the QSAR model was compared to the RBA of estradiol, the positive control used in deriving the RBA of the ER QSAR model. The RBA for compounds was calculated relative to the AC_{50} value of estradiol and converted into a percentile. Comparison of RBA levels for *in silico* vs. *in vitro* was a poorer match for the High and Medium binding levels of the human receptor compared to the other two species. This was observed for both in-domain and total compounds, as most of the compounds that weren't predicted were in the 'None' category (Table 4). For AR binding, the potency-based prediction wasn't performed as we were unable to obtain AC_{50} values for R1881, the positive control used in deriving RBA of AR QSAR model. The AC_{50} values were not available in the raw data files for R1881. Correspondence with EPA suggested that the R1881 was used as a positive control at a couple of concentrations only and not the full concentration response. Also, the AC_{50} wasn't calculated based on the potency performance of the positive control (e.g., relative binding) but instead, the positive control wells were normalized to derive response values (efficacy) that were modeled across the tested concentration range for the entire ToxCast chemical library. The Office of Prevention, Pesticides and Toxic Substances (OPPTS) EDSP test guideline for conducting the AR binding assay using rat prostate cytosol suggests blocking the potential interaction of the ligand with Progesterone

receptors as part of the assay procedure. Based on personal communication with EPA, the procedures of the NVS assay didn't address such matters.

Probability of positive and negative predictions. The value of the QSAR findings for improving the identification of ER and AR binding was determined using Bayesian analyses. The fraction of the tested compounds in ToxCast II was used as the 'prior' and the 'posterior' probabilities of being positive or negative in the assay for all compounds as well as for those in the domain of the respective model. The probabilities were determined for each of the three ER and AR binding assays (Table 5). For ER, the value of the QSAR was greatest for the in-domain compounds, where a positive QSAR prediction indicated a 51% chance of being positive in the human ER assay and a negative finding indicated a 96% chance of being negative in the human ER assay. For AR, the value of the QSAR was greatest for the in-domain compounds, where a positive QSAR prediction indicated a 52% chance of being positive in the human AR assay and a negative finding indicated a >99% chance of being negative in the human AR assay.

Assessment of the QSAR models and the assays used to derive them is also crucial to understand the internal predictivity of the model and for comparison to the ToxCast assays. The following sections deal with the predictive ability of the models for the ToxCast compounds, comparison of the assays used to derive the QSAR models with the ToxCast assays, and the internal predictivity of the QSAR models.

Estrogen Receptor (ER) QSAR prediction for ToxCast compounds. Out of 1851 ToxCast compounds having *in vitro* ER assay AC₅₀ data for all mammalian NR targets, 6 compounds were not predicted by the ER QSAR model out of which 5 were not-active in two or more than two assays and one (Fulvestrant) was active in all three ER assays. For the remaining 1845

compounds that were predicted, 1680 (91%) were not-active in any of the three ToxCast *in vitro* assays but 74 were predicted active by the *in silico* model. Out of these 74, 41 were in the training set of the model with existing experimental data, among them 9 were not-active and 32 compounds were active with a different range of RBA activity (greater than 0% to maximum 10%). Similarly, out of 1845 compounds, 45 (2.5%) were active in all three ToxCast assays out of which 28 were predicted active (20 in the training set of the ER model) and 17 were predicted not-active (none in training set) by the *in silico* model. The remaining 120 compounds were active in some assays and inactive in others.

Androgen Receptor (AR) QSAR prediction for ToxCast compounds. Out of 1851 ToxCast Phase II compounds for AR, 93 were not predicted by the AR QSAR model (undefined in ranges $0.001 < \text{RBA} < 0.1\%$ and with missing parameter), out of which 66 were not-active and 5 were active in all three AR *in vitro* assays and the remaining 22 had different assay results. For the remaining 1758 compounds that were predicted, 800 (45.5%) didn't have any value for the chimp AR binding assay (NVS_NR_cAR) and excluding those, 775 (44.1%) were not-active in any of the three ToxCast *in vitro* assays, but 74 of them were predicted active by the *in silico* model. Out of these, 10 were in the training set of the model with experimental *in vivo* data, where 1 was not-active and 9 were active with a different range of RBA activity (minimum 0.001% to maximum 0.1%). Similarly, 39 (2.2%) compounds were active in all three ToxCast assays out of which 20 were predicted active, 14 were predicted not-active and 5 were not-predicted by the *in silico* model.

Comparison of the ER assay used for the QSAR Model vs. ToxCast ER binding assays. Restricting the analysis to compounds with relatively high *in vitro* activity, i.e., $\text{RBA} > 0.1\%$ that were in the ER QSAR training set (17 compounds), we found 14 to be active in all three ToxCast

in vitro assays of mouse, bovine and human. Out of these, 4-nonylphenol (linear) was inactive in all three ToxCast *in vitro* binding assays (but active in 10 out of 11 of the remaining agonist-mode assays), 4-dodecylphenol was inactive in 2 out of 3 assays (bovine and mouse) and mestranol was inactive only in 1 (mouse) out of 3 ToxCast assays (see Figure 4). Similarly, for inactive compounds present in the ER QSAR model (107 compounds), 1, 9 and 7 were active in the bovine, human and mouse ToxCast assays, respectively. Phenolphthalein was active in all three ToxCast assays whereas Phenol red, 2,2',6,6'-Tetrachlorobisphenol A and 4-(Hexyloxy)phenol were active in 2 (human and mouse) out of 3 assays. There were instances of compounds such as phthalates which were considered active in the training set for the ER QSAR model but inactive in the ToxCast *in vitro* assay (Supplemental Material, Table 3c). This highlights the contrary assay results for the same compounds between different *in vitro* assay platforms in comparison to ToxCast.

Comparison of the AR assay used for the QSAR Model vs. ToxCast AR binding assays. Restricting the analysis to compounds with relatively high activity (RBA >0.1%) that are in the AR QSAR training set (46 compounds), only 2 (Cyproterone acetate and Hydroxyflutamide) were active in all three ToxCast *in vitro* assays of rat, chimp and human. Similarly, there were 30 compounds that were inactive in the training set of the model, of which all were inactive or not predicted (NA) in all three ToxCast *in vitro* assays.

QSAR model internal predictivity of training set compounds. There were cases when some of the training set compounds used to derive the model were either positive in the *in vitro* ER binding assay but the QSAR model predicted no binding activity (Supplemental Material, Table 4a) or negative *in vitro* but predicted to be positive by the QSAR (Supplemental Material, Table 4b). Similarly, there were cases when some of the compounds were either positive in the AR assay

and also belonged to the training set of the model but their predictions were predicted negative by QSAR (Supplemental Material, Table 5a) or were negative *in vitro* but predicted to be positive by the QSAR (Supplemental Material, Table 5b). Out of 190 compounds with relatively high RBA activity in the ER QSAR model, 31 were predicted to be out of the total applicability domain of the model. Similarly, out of 76 compounds with relatively high RBA activity of the AR QSAR model, 19 were predicted to be out-of the total applicability domain of the model. The full list of compounds with their names, SMILES codes, *in vitro* assay values and *in silico* predictions is given as excel file, Supplemental Material, Excel Table 2a.

Discussion

The OASIS 3D QSAR models that were used for estimating ToxCast *in vitro* data do not have 100% predictive capability. There are active compounds that belonged to the training set of the model and were in-domain but the QSAR model predicted no binding activity. Moreover, some of the compounds were predicted opposite in activity by the ER or AR QSAR models. This lack of predictivity or opposite prediction could be due to an undefined RBA activity or some missing parameters needed for the prediction. This reduces the prediction performance of the models when using them to further evaluate novel compounds. We also noticed contrary assay results for the same compounds found in the *in silico* model and ToxCast *in vitro* data. Because the *in vitro* assay data platform used to develop the OASIS-QSAR model is different than that of the ToxCast assay we expected to find subtle differences in assay activities and predictions especially for weaker binders. The differences in assay activities were found more in the ER platform than the AR platform for the same compounds.

The sensitivity and specificity for the *in silico* models for binding assays were above 80% on average for the in-domain compounds. This reveals a robust and predictive QSAR model developed on multiple assays and species (*in vitro* human and trout data) platforms can be used to predict other *in vitro* and/or *in vivo* data generated in different labs, species or assay platforms. When the evaluation was restricted to include only the stronger binders ($AC_{50} < 1 \mu\text{M}$), the prediction of binding for the parents or known hydroxylated metabolites improved the sensitivity of the prediction to 100% for both the ER and AR models. When the activities of this restricted set were further examined for the other *in vitro* ToxCast assays within the ER or AR platforms or for curated uterotrophic data (ER platform), the results showed that a QSAR prediction of receptor binding for compounds in-domain flagged compounds that were always active in the other *in vitro* or *in vivo* screening assays. This suggests a tiered screening approach wherein QSAR is first used to identify compounds in-domain of the ER or AR binding models and predicted to bind, which are then screened *in vitro* to assess binding potency, with the stronger binders ($AC_{50} < 1 \mu\text{M}$) are screened *in vivo*. It is important to emphasize that this approach would only identify the subset of compounds that were in-domain of the QSAR, flagged as potential binders and then shown to bind with strong affinities across three independent platforms. This approach would not apply to compounds that were not in domain (majority of compounds), predicted to not bind, require metabolism to manifest activity, were not active in *in vitro* binding assays, or for the AR platform where the relationship of the QSAR or *in vitro* assays to *in vivo* data was not studied. Because the binding assays do not accommodate metabolism, consideration should be given to simulating this within the initial QSAR model analysis. For compounds predicted by QSAR to only bind after being metabolized, any negative binders would need to be followed up by other screening approaches that include metabolism.

The ToxCast assays can be a robust source of data to improve the existing model predictions or derive novel *in silico* models with improved predictivity or refine the existing QSAR model. For example, in the case of Phthalates which are known to be ER inactive (Moore 2000) (see uterotrophic assay data as well as ToxCast *in vitro* binding data) the data used in the derivation of the TIMES ER QSAR model considered them as active compounds. Phthalates undergo ester hydrolysis *in vivo* which might explain the discrepancy between *in vitro* and *in vivo* assay results. Interestingly, compounds like butyl benzyl phthalate (BBP) (Picard et al. 2001) are active in almost all ToxCast ER assays except the NVS ER binding assays. Further analyses are needed to address quantitative prediction of low or medium levels of ER and AR binding using *in silico* approaches. The differences in ER and AR binding activity between species (human, bovine, rat, mouse, chimp, etc.) need to be further explored.

Conclusions

We assessed the OASIS 3D-QSAR models for predicting ER and AR binding by using respective *in vitro* HTS binding data from >1800 ToxCast Phase II compounds generated by NVS. Our analysis indicated that for ER, the QSAR predictions of the three NVS assay platforms' results had low sensitivity (< 56%) but high specificity (95%) and concordance (>91%) when all compounds in the dataset were analyzed. For the in-domain compounds, the ER QSAR model had high sensitivity (>75%) high specificity (>86%) and overall concordance decreased by approximately 5%. When HT results were restricted to a subset of compounds within the domain of the ER QSAR model and with consistent agreement of ER binding at $AC_{50} < 1 \mu\text{M}$ for the three binding assays, the ER QSAR model predicted binding for the parents or known hydroxylated metabolites 100% of the time. Similarly, for AR, QSAR predictions of the three assay platform results had low sensitivity (<41%) but high specificity (84-89%) and

concordance (>83%) when all compounds in the dataset were analyzed. For the in-domain compounds the AR QSAR model had high sensitivity (92-100%), specificity of 70-81% and overall concordance decreased by approximately 10%. Similarly, when HT results were restricted to a subset of compounds within domain of the AR QSAR model and with consistent agreement of AR binding at $AC_{50} < 1 \mu\text{M}$ for the three binding assays, the QSAR model accurately predicted binding for the parent compounds 100% of the time.

The potency of binding prediction for *in vitro* compounds was estimated only for the ER model where *in silico vs. in vitro* comparison was found to be a poorer match for High and Medium RBA levels for the *in vitro* human receptor, compared to receptors for the other two species.

Heatmaps showed that the subset of ToxCast compounds that were in domain of the ER or AR QSAR models and predicted to be active were active across all ER and AR binding platforms ($AC_{50} < 1 \mu\text{M}$). This subset of compounds was also active in the respective transactivation assay where metabolism was incorporated with addition of exogenous S9 fraction (OT platform). For those compounds in this subset known to require metabolism to manifest activity, the transactivation response appears to be less promiscuous than binding because of the mixed nature of the response in these assays, which may reflect some degree of constitutive metabolism depending on the cell type. This same subset of ER active compounds was active in the uterotrophic assay *in vivo*. Based on this research, a tiered screening approach could be implemented wherein (1) QSAR is used to identify compounds in-domain of the ER or AR binding models and predicted to bind; (2) those compounds are screened *in vitro* to assess binding potency; and (3) the stronger binders ($AC_{50} < 1 \mu\text{M}$) are screened *in vivo*. It is important to emphasize that this approach would only identify the subset of compounds that were in-domain of the QSAR, flagged as potential binders and then shown to bind with strong affinities

across three independent platforms. This approach would not apply to compounds that were not in-domain, predicted to not bind, require metabolism to manifest activity, were not active in *in vitro* binding assays, or for the AR platform where the relationship of the QSAR or *in vitro* assays to *in vivo* data was not studied. In such scenario mathematical models using battery of *in vitro* assays could be useful.

References:

- Bradbury S, Kamenska V, Schmieder P, Ankley G, Mekenyan O. 2000. A computationally based identification algorithm for estrogen receptor ligands: Part 1. Predicting her α binding affinity. *Toxicol Sci* 58:253-269.
- Browne P, Judson RS, Casey W, Kleinstreuer N, Thomas RS. 2015. Screening chemicals for estrogen receptor bioactivity using a computational model. *Environ Sci Technol* 49:8804-8814.
- Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. 2007. The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* 95:5-12.
- U.S. EPA EDSP21 dashboard. <http://actor.Epa.Gov/edsp21/>.(accessed 12/31/2015)
- Fang H, Tong W, Branham WS, Moland CL, Dial SL, Hong H, et al. 2003. Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor. *Chem Res Toxicol* 16:1338-1358.
- Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, et al. 2015. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high throughput screening assays for the estrogen receptor. *Toxicol Sci*.
- Katzenellenbogen JA. 1995. The structural pervasiveness of estrogenic activity. *Environ Health Perspect* 103 Suppl 7:99-101.
- Katzenellenbogen JA, Muthyala R, Katzenellenbogen BS. 2003. The nature of the ligand-binding pocket of estrogen receptor alpha and beta: The search for subtype-selective ligands and implications for the prediction of estrogenic activity. *Pure Appl Chem* 75, 2397-2403.
- Kavlock R, Chandler K, Houck K, Hunter S, Judson R, Kleinstreuer N, et al. 2012. Update on epa's toxcast program: Providing high throughput decision support tools for chemical risk management. *Chem Res Toxicol* 25:1287-1302.
- Kelce WR, Monosson E, Gamcsik MP, Laws SC, Gray LE, Jr. 1994. Environmental hormone disruptors: Evidence that vinclozolin developmental toxicity is mediated by antiandrogenic metabolites. *Toxicol Appl Pharmacol* 126:276-285.
- Kolsek K, Mavri J, Sollner Dolenc M, Gobec S, Turk S. 2014. Endocrine disruptome--an open source prediction tool for assessing endocrine disruption potential through nuclear receptor binding. *J Chem Inf Model* 54:1254-1267.
- McRobb FM, Kufareva I, Abagyan R. 2014. In silico identification and pharmacological evaluation of novel endocrine disrupting chemicals that act via the ligand-binding domain of the estrogen receptor alpha. *Toxicol Sci* 141:188-197.
- Mekenyan O, Ivanov J, Karabunarliev S, Bradbury SP, Ankley GT, Karcher W. 1997. A computationally-based hazard identification algorithm that incorporates ligand flexibility. 1. Identification of potential androgen receptor ligands. *Environ Sci Technol* 31:3702-3711.

- Mekenyan OG, Kamenska V, Schmieder PK, Ankley GT, Bradbury SP. 2000. A computationally based identification algorithm for estrogen receptor ligands: Part 2. Evaluation of a heralpha binding affinity model. *Toxicol Sci* 58:270-281.
- Moore NP. 2000. The oestrogenic potential of the phthalate esters. *Reprod Toxicol* 14:183-192.
- Picard K, Lhuguenot JC, Lavier-Canivenc MC, Chagnon MC. 2001. Estrogenic activity and metabolism of n-butyl benzyl phthalate in vitro: Identification of the active molecule(s). *Toxicol Appl Pharmacol* 172:108-118.
- Piparo EL, Worth A. 2010. Review of qsar models and software tools for predicting developmental and reproductive toxicity. Ispra (VA), Italy:Joint Research Centre. Available: <http://publications.jrc.ec.europa.eu/repository/handle/JRC59820> (accessed 07/27/2015).
- Ravdin PM, van Beurden M, Jordan VC. 1987. Estrogenic effects of phenolphthalein on human breast cancer cells in vitro. *Breast Cancer Res Treat* 9:151-154.
- Serafimova R, Todorov M, Nedelcheva D, Pavlov T, Akahori Y, Nakai M, et al. 2007. Qsar and mechanistic interpretation of estrogen receptor binding. *SAR QSAR Environ Res* 18:389-421.
- Shelby MD, Newbold RR, Tully DB, Chae K, Davis VL. 1996. Assessing environmental chemicals for estrogenicity using a combination of in vitro and in vivo assays. *Environ Health Perspect* 104:1296-1300.
- TIBCO SpotFire Spotfire for data visualization and decision making, Available: <http://spotfire.tibco.com/>. (accessed 12/31/2015)
- Steinmetz FP, Mellor CL, Meini T, Cronin MTD. 2015. Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: Using public data to build screening tools within a knime workflow. *Mol Informatics* 34:171-178.
- Waller CL, Juma BW, Gray LE, Jr., Kelce WR. 1996. Three-dimensional quantitative structure--activity relationships for androgen receptor ligands. *Toxicol Appl Pharmacol* 137:219-227.
- Zhang L, Sedykh A, Tripathi A, Zhu H, Afantitis A, Mouchlis VD, et al. 2013. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using qsar- and structure-based virtual screening approaches. *Toxicol Appl Pharmacol* 272:67-76.

Tables

ER binding	Human			Bovine			Mouse		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Positive	44	13	57	31	3	34	31	10	41
Negative	43	358	401	56	368	424	56	361	417
Total	87	371	458	87	371	458	87	371	458
Sensitivity (%)	44/57 = 77.2			31/34 = 91.2			31/41 = 75.6		
Specificity (%)	358/401 = 89.3			368/424 = 86.8			361/417 = 86.6		
Concordance (%)	(44+358)/458 = 87.8			(31+368)/458 = 87.1			(31+361)/458 = 85.6		
Coverage	[(44+358)/1845]*100 = 21.8%			[(31+368)/1845]*100 = 21.6%			[(31+361)/1845]*100 = 21.2%		
AR binding	Human			Chimp			Rat		
	Positive	Negative	Total	Positive	Negative	Total	Positive	Negative	Total
Positive	36	2	38	25	0	25	23	2	25
Negative	33	142	175	32	77	109	46	142	188
Total	69	144	213	57	77	134	69	144	213
Sensitivity (%)	36/38 = 94.7			25/25 = 100			23/25 = 92		
Specificity (%)	142/175 = 81.1			77/109 = 70.6			142/188 = 75.5		
Concordance (%)	(36+142)/213 = 83.6			(25+77)/134 = 76.1			(23+142)/213 = 77.5		
Coverage	[(36+142)/1758]*100 = 10.1%			[(25+77)/958]*100 = 10.6%			[(23+142)/1758]*100 = 9.38%		

Table 1. Summary performance of QSAR model predictions for ToxCast II compounds against individual mammalian *in vitro* assays 458 in-domain compounds for Estrogen Receptor (ER) binding model v.03 (top) and 213 (134 for chimp) in-domain compounds for Androgen Receptor (AR) binding model v.03 (bottom).

Compound Name	Predicted result	Total Domain	NVS_NR_bER	NVS_NR_hER	NVS_NR_mERa
Clomiphene citrate*	Not Active	Out of Domain	0.0317	0.00975	0.187
17alpha-Estradiol	Active	In domain	0.000493	0.0000595	0.0229
17beta-Estradiol	Active	In domain	0.000174	0.0229	0.00164
4-Hydroxytamoxifen	Active	In domain	0.00186	0.0025	0.0723
Diethylstilbestrol	Active	In domain	0.0229	0.0229	0.00632
17alpha-Ethinylestradiol	Active	In domain	0.000245	0.0000541	0.00185
Bisphenol A	Active	In domain	0.389	0.131	0.15
Bisphenol AF	Active	In domain	0.096	0.0449	0.0242
Bisphenol B	Active	In domain	0.149	0.0291	0.022
2,2',4,4'-Tetrahydroxybenzophenone	Active	In domain	0.268	0.0534	0.176
Daidzein	Active	In domain	0.481	0.116	0.173
Estriol	Active	In domain	0.00763	0.0229	0.0421
Estrone	Active	In domain	0.104	0.000795	0.00763
meso-Hexestrol	Active	In domain	0.000277	0.0229	0.0229
HPTE	Active	In domain	0.0176	0.0392	0.00985
Genistein	Active	In domain	0.0983	0.0167	0.0901
Raloxifene hydrochloride*	Not Active	Out of Domain	0.00763	0.0000476	0.0253
Phenolphthalein	Active	Out of Domain	0.658	0.887	0.228
Tamoxifen	Not Active	Out of Domain	0.0834	0.0349	0.133
Tamoxifen citrate*	Not Active	Out of Domain	0.106	0.0246	0.223

Table 2. Twenty compounds that have ER binding at $AC_{50} < 1 \mu\text{M}$ for all three mammalian nuclear receptor binding assays. The *in silico* prediction results including the total domain information as well as *in vitro* assay data are given. *The salt or the acid component is removed for QSAR modeling.

Compound Name	Predicted result	Total Domain	NVS_NR_cAR	NVS_NR_hAR	NVS_NR_rAR
17alpha-Estradiol	Active	In domain	0.024	0.0057	0.242
17beta-Estradiol	Active	In domain	0.0167	0.00293	0.12
17beta-Trenbolone	Active	In domain	0.00744	0.000201	0.0179
17-Methyltestosterone	Active	In domain	0.00614	0.00144	0.0802
5alpha-Dihydrotestosterone	Active	In domain	0.00763	0.000566	0.022
Cyproterone acetate*	Active	In domain (belongs to training set)	0.0326	0.00763	0.258
Mifepristone	Can't Predict	Out of Domain	0.0388	0.0282	0.0621
Norethindrone	Active	In domain	0.00396	0.000505	0.147
Norgestrel	Active	In domain	0.00538	0.00131	0.093
Progesterone	Active	In domain	0.163	0.00763	0.414
Spirolactone	Active	In domain	0.0136	0.00303	0.254

Table 3. Eleven compounds that have AR binding at $AC_{50} < 1 \mu\text{M}$ for all the three mammalian nuclear receptor binding assays. The *in silico* prediction results including the total domain information as well as *in vitro* assay data are given. *The salt or the acid component is removed for QSAR modeling.

In Domain <i>in vitro</i>	Bovine						Human						Mouse					
	<i>in silico</i>					Total	<i>in silico</i>					Total	<i>in silico</i>					Total
	Hi	Med	Low	VLow	None		Hi	Med	Low	VLow	None		Hi	Med	Low	VLow	None	
Hi	3	1	<i>0</i>	0	0	4	8	7	7	0	2	24	3	2	0	0	0	5
Med	5	6	<i>1</i>	0	0	12	1	6	<i>11</i>	4	7	29	4	3	8	0	1	16
Low	0	1	5	0	1	7	0	0	0	0	4	4	0	4	3	1	10	18
VLow	0	4	3	2	2	11	0	0	0	0	0	0	0	0	0	0	2	2
Total	8	12	9	2	3	34	9	13	18	4	13	57	7	9	11	1	13	41

Table 4. Summary performance of QSAR model potency predictions for in-domain ToxCast II compounds against individual mammalian *in vitro* assays for Estrogen Receptor (ER) binding model v.03. Here, data are **bolded** to show agreement and *italicized* and **bolded** to show disagreement. Also, Hi = high, Med = medium, and VLow = very low. Potency bins are described in the methods section “Performance of QSAR models”.

	Estrogen Receptor						Androgen Receptor					
	ToxCast (All)			ToxCast (In-domain)			ToxCast (All)			ToxCast (In-domain)		
	Cells			Cells			Cells			Cells		
	Human	Bovine	Mouse	Human	Bovine	Mouse	Human	Chimp	Rat	Human	Chimp	Rat
Probability of positive in ToxCast dataset (%)	6.7	3.5	5.6	12	7.4	9	7	10	6	18	19	12
Probability of positive if QSAR is positive (%)	41	28	30	51	36	36	25	26	17	52	44	33
Probability of negative in ToxCast dataset (%)	93.3	96.5	94.4	88	92.6	91	93	80	94	82	81	88
Probability of negative if QSAR is negative (%)	96	98	96	96	99	97	95	92	95	99	100	99

Table 5. Value of QSAR findings in improving the prediction of ER binding (left) and AR binding (right).

Figure legends

Figure 1: Tamoxifen, Clomiphene (which are triphenylethylenes) and Raloxifene (which is a benzothiophene) are common ER binders used in clinical practice for treatment of breast cancer, induction of ovulation in sub-fertile women, and prevention of post-menopausal osteoporosis respectively. Phenolphthalein, used in nonprescription laxative preparations, also has a weak estrogenic action (Ravdin et al. 1987). The site of 4-hydroxylation for Tamoxifen and Clomiphene is shown with an asterisk.

Figure 2: Heatmaps of 18 ER (left) and 11 AR (right) compounds with $AC_{50} < 1\mu\text{M}$ (most active) for ToxCast assays using TIBCO Spotfire. Color codes indicate least to most active compounds with the increasing gradient of brown color, darker shade indicates more potent activity (lower AC_{50}) and black represents inactive compounds.

Figure 3: Progesterone and its synthetic analogs which are progesterone receptor binders and in higher doses can bind AR. They are used in clinical practice to induce abortion, treat premenstrual syndrome and pain, as hormonal contraceptives, and to reduce elevated or unwanted androgen activity in the body respectively.

Figure 4: Representative compounds for which *in vivo* and *in vitro* assay results are not concordant. These compounds were present in the training set of the ER QSAR model and the data were compiled from *in vivo* studies (Serafimova et al. 2007). The first three compounds were active *in vivo* but inactive *in vitro* whereas the latter three were active *in vitro* but inactive *in vivo*.

Figure 1.

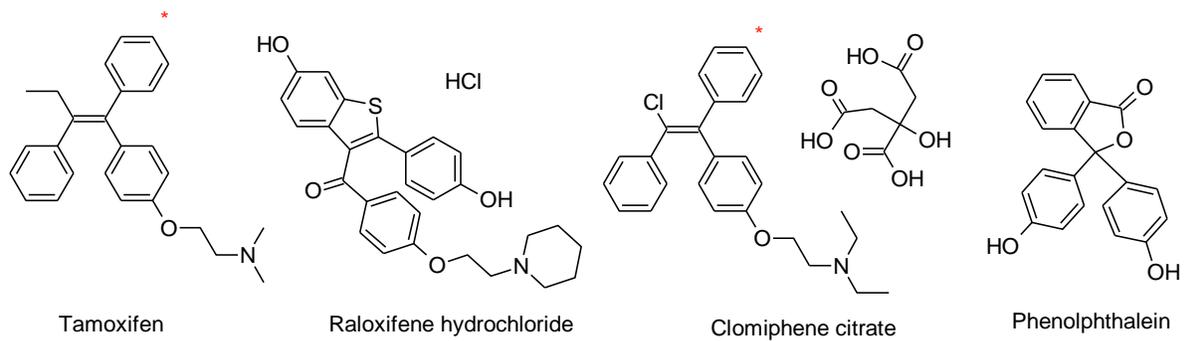


Figure 2.

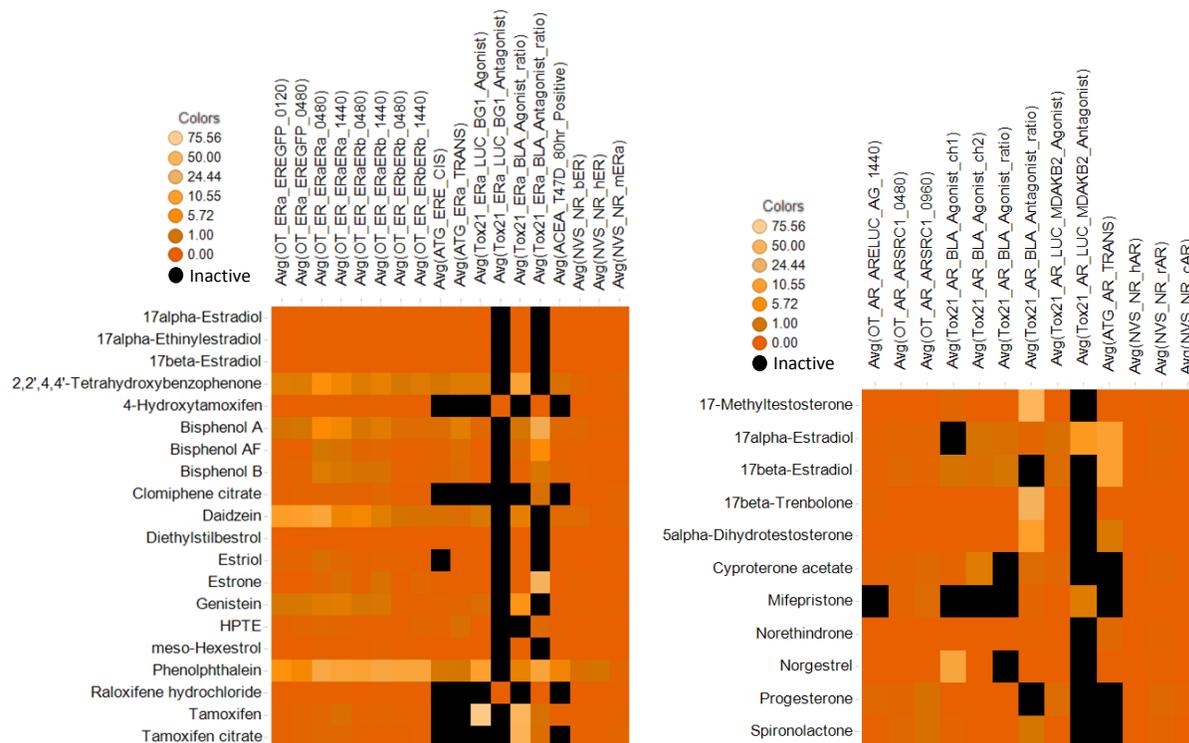


Figure 3.

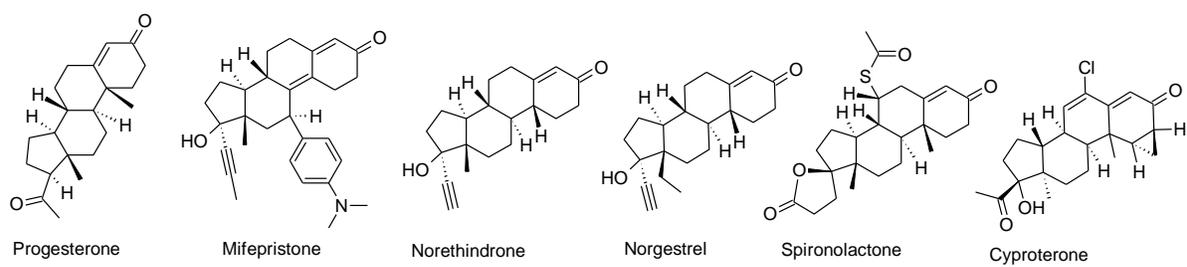


Figure 4.

