HYBRID MODELING OF ESTROGEN RECEPTOR BINDING AGENTS USING ADVANCED CHEMINFORMATICS TOOLS AND MASSIVE PUBLIC DATA

by

KATHRYN RIBAY

A thesis submitted to the

Graduate School-Camden

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Chemistry

written under the direction of

Dr. Hao Zhu

and approved by

Dr. Hao Zhu

Dr. Joseph Martin

Dr. Alex Roche

Dr. Georgia Arbuckle-Keil, Graduate Program Director

Camden, New Jersey

January 2016

THESIS ABSTRACT

Hybrid Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data

by KATHRYN RIBAY

Thesis Director:
Dr. Hao Zhu

Estrogen receptor-α (ERα) is a critical target for drug design as well as a potential source of toxicity when activated unintentionally. Thus, evaluating potential ERα binding agents is critical in both drug discovery and chemical toxicity areas. Using computational tools, e.g. Quantitative Structure-Activity Relationship (QSAR) models, can predict potential ERα binding agents before chemical synthesis. The purpose of this project was to develop enhanced predictive models of ERα binding agents by utilizing advanced cheminformatics tools that can integrate publicly available bioassay data. The initial ERα binding agent data set, consisting of 446 binders and 8,307 non-binders, was obtained from the Tox21 Challenge project organized by the NIH Chemical Genomics Center (NCGC). After removing the duplicates and inorganic compounds, this data set was used to create a training set (259 binders and 259 non-binders). This training set was used to develop QSAR models using chemical descriptors. The resulting models were then used to predict the binding activity of 264 external compounds, which were available to us after the models were developed. The cross-validation results of training set [Correct Classification Rate (CCR)) = 0.72] were much higher than the external predictivity of the unknown compounds (CCR= 0.59). To improve the conventional QSAR models, all compounds in the training set were used to search PubChem and generate a profile of their biological responses across thousands of bioassays. The most important bioassays were prioritized to generate a similarity index that was used to calculate the biosimilarity score between each two compounds. The nearest neighbors for each compound within the set were then identified and its ERα binding potential was predicted by its nearest neighbors in the training set. The hybrid model performance (CCR=0.94 for cross validation; CCR=0.68 for external prediction) showed significant improvement over the original QSAR models, particularly for the activity cliffs that induce prediction errors in conventional QSAR models. The results of this study indicate that the response profile of chemicals from public data provides useful information for modeling and evaluation purposes. The public big data resources should be considered along with chemical structure information when predicting new compounds, such as unknown ERα binding agents.

ACKNOWLEDGEMENTS

I would first and foremost like to thank my mentor, Dr. Hao Zhu for convincing me to switch to the research thesis track as well as for his constant encouragement throughout the process. I am so grateful for all that I have learned in his lab, and I appreciate his patience as well as his genuine commitment to developing the abilities of his students.

I would also like to thank my committee members, Dr. Joseph Martin and Dr. Alex Roche for taking the time to advise me as well. Thank you to Dr. Georgia Arbuckle for your guidance.

Finally, I would like to thank Marlene Kim for her unending patience and willingness to help me. I would also like to thank my other lab mates (Wenyi Wang, Dan Russo, Dan Pinolini) for their help and suggestions.

DEDICATION

I would like to dedicate this thesis to my husband, Randy Ribay, who now knows far more about computational chemistry than the average English teacher.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND RESEARCH GOALS

Section 1: Estrogen Receptors and Estrogen Disrupting Chemicals

Estrogen receptors (ER) are cellular proteins that are activated when bound to estrogen molecules. When activated, estrogen receptors trigger the expression of gene products crucial to the endocrine system.[1] ER are members of the broader family of nuclear receptors. These receptors, which include receptors for steroid hormones, thyroid hormones, vitamin D and retinoids, They are characterized by a small DNA binding domain and a large ligand binding domain at the C-terminal end of the protein.[2] There are two unique estrogen receptors: ERα and ERβ. ERβ was first identified in 1992 in rat prostate and ovary tissue, while ERα has been characterized for much longer.[3] These two receptors are highly similar in the DNA binding domain, which binds to the estrogen response elements in target genes, but differ more significantly in other regions. While the DNA binding domain is 97% homologous between the two receptors, the ligand binding domain of ERβ is only 60% homologous to that of ERα.[4]

In the most straightforward estrogen receptor interactions, the estrogen compound binds to the ligand binding domain of the estrogen receptor and the resulting complex recruits the required coregulators to interact with the estrogen response elements in the target genes and carry out the targeted gene transcription.[4] The estrogen receptor binds as a dimer in the presence of an agonist ligand.[5] This process can become increasingly complex depending on the agonist or antagonist nature of the binding agent, the cofactors, and the targeted gene.[4,6-9] For both ERα and ERβ, the primary agonist is 17-β estradiol (**Figure 1**).[10]

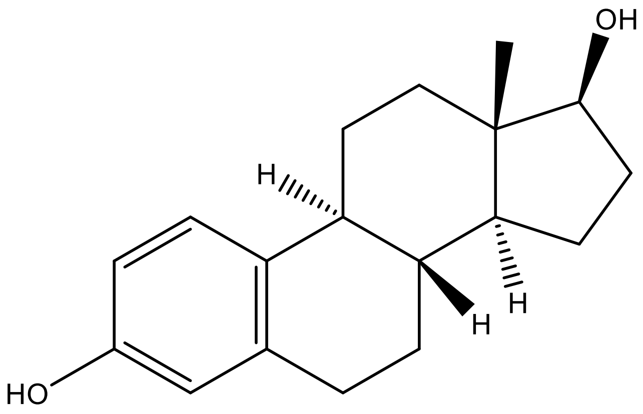


Figure 1: 17-β estradiol

Previous studies have identified the substructural elements strongly associated with ERα binding (biophores) as well as substructural elements strongly associated with nonbinding (biophobes).[11,12] The existence of these fragments and the distance separating them can be used to suggest compounds for future study, but identifying biophore fragments in compounds of known activity does not always correspond to ERα binding in compounds of unknown activity. For example, a phenolic –OH group (**Figure 2**) is a common structural element in ERα binders, yet if the surrounding compound creates steric hindrance that prevents H-bonding, the compound will be inactive.[12]

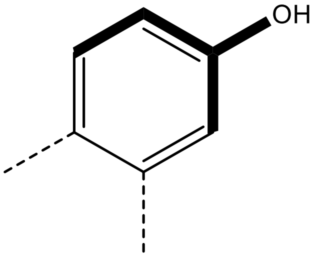


Figure 2: Phenolic –OH group, a substructural element commonly found in ERα binders

Estrogen receptors can also be activated by certain endocrine disrupting chemicals (EDC), resulting in a disruption of normal estrogen signaling. This can lead to a variety of neurological, metabolic, reproductive, and developmental effects since estrogen receptors mediate estrogen signaling in reproductive tissue as well as in the brain, lungs, and cardiovascular system.[6] Like hormones, many EDC can have tissue-specific action, even at low doses. The disruption is especially harmful to children and developing fetuses.[7] In some pharmaceutical instances that tissue-specific action of ER binding agents has been harnessed, such as in the case of Tamoxifen. Tamoxifen, a selective estrogen receptor modulator, is an ERα antagonist that is used to target ER positive breast cancer tumors.[13] Due to the large ligand bind domain of both estrogen receptors, there are many small molecules which exhibit various degrees of interaction with the receptor. Once bound with the EDC, the ER-ligand complex can cause either agonist or antagonist activity depending on the resulting conformation.[7] While there are many EDC that interact with both receptors, the difference between these two receptors allows some ligands specifically bind to only one receptor. Among all known binding agents, the ERα binders are much better characterized than ERβ binders.[1,6,14] Due to the nature of available data, this study focused solely on ligands binding to ERα.

When estrogen receptors are activated by small molecules other than estrogens, the expression of the associated genes is deregulated leading to neurological, developmental, and reproductive toxicity.[15] When considering the large amount of compounds which need to be evaluated for their estrogen receptor binding potentials, traditional experimental toxicology protocols can be costly and time-consuming. As a result, there is a strong need to effectively pre-screen and prioritize small molecules for potential endocrine disruption prior to more costly animal testing. In a 2007 publication, the U.S. National Research Council identified both high-throughput screening (HTS) and computational models as critical chemical toxicity evaluation tools in 21st century toxicology.[16] HTS, which uses an automated robotic systems to conduct a large volume of microplate assay tests rapidly, has been viewed as a potential alternative to animal models due to the ability to test many molecules at a rapid pace and lower cost. The large number of HTS studies has resulted in publically available bioassay databases which are a rich source of *in vitro* data.[17] A downside to HTS is the frequent rate of false positives and false negatives that can occur due to the set concentration which may not correlate with the AC_{50} values of the tested compound.[18] Quantitative high-throughput screening (qHTS) builds on HTS by incorporating the ability to test varying compound concentrations in order to provide more detailed biological activity data. qHTS. This allows for more useful information that can be used to more confidently assess the activity of small molecule compounds. Motivated by these available data, computational modeling, which costs even less than HTS or qHTS, has gained attention as another important evaluation protocol for EDC.[19]

Section 2: QSAR Modeling Studies

Quantitative structure-activity relationship (QSAR) modeling was originally developed to identify lead compounds in drug design but has gained importance in recent years for wider applications.[20] QSAR modeling uses mathematical equations to define a relationship between the structure of a compound and its biological activity and then uses this quantitative relationship to make predictions of the activity of unknown compounds. The QSAR model development process consists of three steps: data preparation, data analysis/ model preparation, and model validation.[21] Each step in this process contains potential pitfalls that current best practices are designed to avoid. Data preparation requires that the bioassay information used be reliable and have used consistent methodology across all compounds.[22] In most large databases, there exist errors in the recorded chemical structures of the compounds, in some cases as high as 10%.[23] Most databases record the chemical compounds as Simplified Molecular Input Line Entry System (SMILES) strings, and incorrect translation or human entry error can lead to mistakes in the chemical structure[24] and, as a result, the chemical descriptors used to build the QSAR model. Therefore, prior to use in QSAR modeling, the chemical compounds must be screened for errors and erroneous chemical structures should be corrected or discarded. Additionally, it is important to avoid using compounds beyond the scope of the chemical descriptors to be used in the QSAR model as well as remove any duplicates. Compounds can either be associated with an exact value for activity or classified into categories (i.e. active/inactive). Finally, the data set must be large enough to incorporate a diversity of chemical structures that will create a more robust QSAR model.[23-25]

Once the database is checked and curated, the chemical structures are used to generate chemical descriptors. Chemical descriptors can be categorized as one-dimensional (such as atom or bond counts), two dimensional (such as path lengths or connectivity), or three dimensional (spatial properties of molecules). In this project, three-dimensional descriptors were not used. Prior to the development of a QSAR model, all descriptors must be normalized to the same [0-1] range scale, and redundant descriptors are removed. Outliers can create be identified by performing a Principal Component Analysis, in which the chemical descriptors are reduced to principal components, and the chemical compounds are mapped onto the most orthogonal principal components.[26] The resulting graph will visually demonstrate if any compounds are significant structural outliers, and may negatively affect the QSAR model development. These compounds should be discarded in order to develop a more effective QSAR model. In addition, it can be used to confirm that the active and inactive compounds are occupying a similar region of chemical space, and will produce a balanced model. However, this process does not allow for the identification of activity outliers, which can create activity cliffs.[27]

QSAR models use a variety of machine-learning methods to develop and test predictions for unknown compounds. In this study, the methods chosen were Support Vector Machines (SVM), Random Forest (RF), and *k*-nearest neighbor (*k*NN). These modeling methods are discussed in further detail in the modeling methods section of Chapter 2. In order to validate the models, both cross-validation and external prediction are used. Five-fold cross validation consists of dividing the training set into five equal subsets. Each subset is removed once, while the remaining four subsets are used to develop the QSAR models. The resulting models are then used to predict the activity of the withheld subset.[28] Using this method, each compound in the training set is predicted once. These predictions are compared to the actual activity in order to evaluate the QSAR model.

While cross-validation is important in order to verify that the training set has produced a model that is consistent across its internal chemical space, frequent studies have shown that external prediction often shows lower accuracy than cross-validation.[22,29] Thus, it is essential to use an external test set to in order demonstrate model predictivity with confidence.[21] For external test set validation, the QSAR model is developed using the full training set. This model is used to predict the biological activity of all compounds in the test set, and the predictions are compared to the known activities in order to evaluate the effectiveness of the QSAR model.

When developing QSAR models, several modeling methods are used in order to generate multiple predictions for each chemical compound. These predictions are then averaged into a consensus prediction. This avoids the need for arbitrary selection of the “best” QSAR model, as the model which best predicted the chosen test set may show slightly different performance with a different set of compounds. The consensus prediction, based on multiple models using multiple descriptor sets, is more reliable and is better used as justification for selecting a compound for further study.[30,31] As the consensus prediction is an average, it must be assigned a designation of active or inactive by rounding the value down to zero (inactive) or up to one (active). In the case of a test set compound that is highly dissimilar from the compounds in the modeling set, an applicability domain[32] may be used to discard the compound from prediction. This prevents an unjustified extrapolation of the defined structure-activity landscape.[33]

QSAR modeling allows for the computational screening of large sets of compounds, which is both cheaper and faster than animal testing or even HTS. In pharmaceutical chemistry, it is particularly useful for determining potential lead compounds as well as excluding compounds that exhibit unfavorable bioactivities. However, existing QSAR studies have shown both the potential and limitations of the current uses of QSAR modeling.[20,34,35] As this method has moved from the domain of pharmaceutical chemistry alone to the wider field of toxicology, certain problems have arisen. Rather than searching for potential leads, computational chemists are searching for potentially harmful compounds, and the challenge of creating global models that accurately predict general toxicity through QSAR continues to elude.[29]

QSAR modeling has been applied to develop estrogen receptor binding models multiple times in recent years. These studies have covered a wide range of modeling approaches and data set sizes (**Table 1**). In the first study, Liu, Papa, and Grammatica[36] combined information from multiple bioassays to refine the identified ERα activity of 108 compounds. They then used ordinary least squares regression and genetic algorithm to develop a QSAR model using chemical structure descriptors of those compounds and predicted the activity of 28 additional compounds. This model showed strong results but used very small training and test sets. Taha et al.[37] specifically modeled the binding of ligands to ERβ using genetic algorithm and multiple least squares regression to build the QSAR models. Their results showed acceptable prediction within the modeling set, but poor prediction for external compounds. Zhang et. al.[38] and Deng et. al.[39] both developed QSAR models to address both ERα and ERβ. In both cases, the QSAR models were better able to predict binding to ERα than ERβ, in part due to the size of the available datasets used to develop the models. A recent study by Zang, Rotroff, and Judson[40] used a large data set of ERα binding activity from the Tox21 project to develop QSAR models using a combination of random forest (RF) support vector machines (SVM) methods. While this modeling study achieved acceptable predictivity, it used only chemical structure data in developing the QSAR models which can leave the results vulnerable to certain specific prediction errors known as activity cliffs.

Year	Receptor Studied	Data Set Size	Method	Reference
2008	α	108	OLS/GA-VSS	[36]
2010	β	119	GA/MLR	[37]
2013	α/β	546α/137β	kNN (STL & MTL)	[38]
2013	α	8147	SVM/RF	[40]
2014	α/β	81	MLR	[39]

Table 1: A sampling of QSAR studies on estrogen receptor interaction.

Although there have been some promising results, there is increasing evidence that the applicability of these models is only limited to certain compounds.[41,42] A fundamental problem with using structure-activity relationships to predict the performance of unknown compounds is the difficulty of predicting the behavior of chemical space beyond that of the known compounds. This is particularly challenging when a small change in biological mechanism can trigger wide-ranging organism-level effects.[29] Compounds with similar structures may show significantly different activities, leading to prediction errors in QSAR models. These pairs of molecules are known as “activity cliffs” in QSAR studies.[27] This term arises from looking at the structure-activity landscape as a graphical surface. When the structure and activity are closely related and change gradually together, the surface is smooth. However, when highly similar molecules exhibit markedly different activity, the surface becomes jagged.[27,33] QSAR models predict the activity of compounds only based on their chemical structure information, but the presence of activity cliffs can lead to unavoidable prediction errors if there is no other information than chemical structures.[43]

The estrogen receptor has been the target of many modeling studies due to the effects of endocrine disruption that occur when a compound present in the environment or in a consumer product activates the receptor. There is a need for methods that can quickly and effectively screen a wide range of chemicals to identify potential EDC before a product is brought to market. The attempt to use QSAR models to fill this need has been hindered by the structural diversity of the estrogen receptor binders and has reached a bottleneck due to the existence of activity cliffs. Although this study focuses on activation of ERα only, there is a wide range of chemical structures that are able to activate this receptor due to its large ligand binding domain.[6] The lack of experimental data, especially for active compounds (ERα binders), has resulted in activity cliffs in QSAR models based solely on chemical structures and limited the applicability of traditional QSAR modeling methods.

Section 3: Research Goals

The goal of this thesis is to address the need for methods to predict the binding of novel chemical compounds to estrogen receptors in the body, particularly ERα. Since compounds that bind to ERα can interfere with the normal function of the endocrine system, causing deleterious effects on multiple systems in the body, novel compounds in pharmaceutical or other industries must undergo costly testing to ensure that they will not interfere with ERα. Cheminformatics tools can be used to screen compounds in a more cost-effective way, but the prevalence of activity cliffs in QSAR models can lead to unreliable results for certain compounds.

In this thesis, I developed and tested enhanced computational models for estrogen receptor binding agents using both QSAR approaches and a biosimilarity search, which is based on publically available bioassay data. This method incorporates both chemical structure data via QSAR descriptors and biological activity data through bioassay responses. Using the resulting hybrid models, the new compounds can be directly predicted for their estrogen receptor binding potential. The incorporation of biosimilarity search based on extra bioassay data can address the activity cliffs issue of QSAR modeling and improve the prediction accuracy for new compounds.

CHAPTER TWO: Quantitative Structure-Activity Relationship Modeling of Estrogen Receptor Binding Affinity

Section 1: Materials and methods

Data curation

The original dataset used in this study was obtained in two parts separately from the National Center for Advancing the Translational Science (NCATS) via the Tox21 Challenge project. The dataset (PubChem assay AID 743077) consisted of the results of the quantitative High Throughput Screening (qHTS) to identify agonists of the ERα signaling pathway by measuring the expression of a beta lactamase reporter gene controlled by an ERα ligand binding domain (ER-LBD) fusion protein.[44] This dataset was used as the training set in the Tox21 Challenge. This original dataset consisted of 8,753 compounds, of which 446 were categorized as active (ERα binders) and 8307 were categorized as inactive (non-binders). The compounds were processed by the CaseUltra® (www.multicase.com) structure checker tool to remove duplicates and inorganics, resulting in 5,647 unique organic compounds (259 actives and 5,388 inactives). Compounds with incorrect structures as well as mixtures were discarded in the process as well. All of the active compounds were selected for the training set and combined with a random selected 259 inactive compounds to produce a balanced training set of 518 compounds. An additional but much smaller set of compounds not included in the original qHTS data was provided by the Tox21 Challenge project as an external test set to validate the resulting models. This external test set of 297 compounds (25 actives and 272 inactives) was also processed by the CaseUltra® structure checker to remove duplicates and inorganics, resulting in 264 unique compounds (24 actives and 240 inactives).

Figure 3: The hybrid modeling workflow

Chemical descriptors

Once the datasets were curated, chemical descriptors were calculated using two commercial descriptor generators. A total of 192 2-D Molecular Operating Environment® (MOE) (www.chemcomp.com) descriptors were generated using MOE version 2013, which include physical properties, atom and bond counts, connectivity and shape indices, adjacency and distance matrix descriptors, etc. Dragon® (www.talete.mi.it/) version 6 was used to generate 1,259 descriptors including constitutional indices, drug-like indices, connectivity indices, functional group counts, etc. MOE descriptors were used as this standardized descriptor set provides a small but widely varied set of descriptors. Dragon descriptors were used as well in order to generate a wider pool of descriptors. The MOE descriptors were used to identify principal components and the training set underwent Principal Component Analysis (PCA) to ensure that all active and inactive compounds occupy intertwining chemical space.

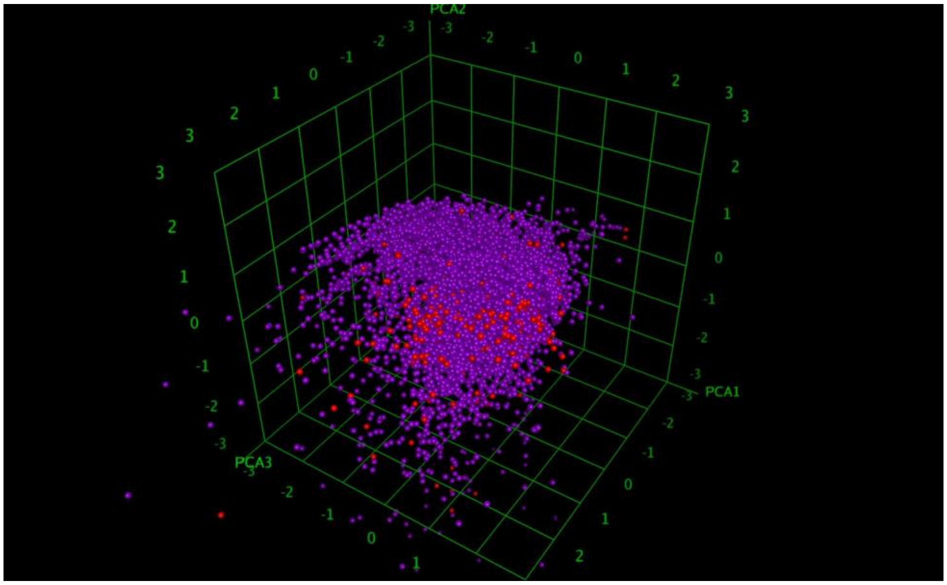


Figure 4: Principal Component Analysis of training set compounds. Red= active; Purple= inactive

All descriptors were normalized to [0,1] and any redundant descriptors were removed by deleting those with low variance (standard deviation <0.01 for the whole training set) and randomly keeping one of any pairs of descriptors that had high correlation (R^2>0.95 between two descriptor values for the training set compounds), leaving 132 unique MOE descriptors and 594 unique Dragon descriptors for both data sets. In order to calculate the chemical similarity among compounds, MOE 2013 was used to calculate 166 MACCS fingerprints of each compound. These fingerprints were used as descriptors to calculate the Tanimoto coefficient of each compound pair to determine their chemical similarity. The Tanimoto coefficient is a measure of the distance between the compounds in chemical space given as a value between [0,1][45]

QSAR model development and model validation

As shown in the hybrid modeling workflow (**Figure 3**), the assay and chemical descriptor data were then used to develop QSAR models via machine learning. Three machine learning algorithms were used to develop QSAR models: support vector machines (SVM), random forest (RF), and *k* nearest neighbor (*k*NN). SVM was originally developed by Vapnik[46] as a method in which modeling set error and model complexity were incorporated into a loss function that was minimized to find the optimal balance of modeling set error and model complexity for predicting the test set. An advantage to using SVM is that it is particularly useful in high-dimensional space, although it becomes less effective if there are significantly more descriptors than compounds.[30] In an SVM QSAR model, the chemical descriptors are mapped onto *n*-dimensional space using kernel functions. The descriptors are mapped for optimal separation between the two different classes of input such that they lie on opposite sides of a dividing hyperplane. The test set compounds are then mapped onto the same space and a prediction is determined based on their location relative to the hyperplane.[46]

For a RF model, the algorithm creates decision trees from randomly selected chemical descriptors. At the start, *n* samples are drawn to create *n* trees. For each data split, *m* variables are randomly selected and the best split from those *m* data points is kept. The trees are not pruned and the process repeats for a defined number of iterations and the outputs are combined for a final prediction. A benefit to RF modeling is that it is efficient for large databases and can detect variable interactions due to the use of bootstrapping during tree growth.[47]

A *k*NN model is based on the concept that compounds will have a similar activity to their nearest neighbors. It uses a classification algorithm method along with a variable selection technique. This model predicts the activity of a target compound by identifying the *k* most similar compounds within the chemical descriptor space and using their activity to predict that of the target compound. A variable selection procedure, in this case genetic algorithm, is used to define the nearest neighbors. The model is originally mapped onto a random selection of chemical descriptors. In genetic algorithm, as in the process of natural selection, descriptors are swapped and the resulting model is evaluated. If the resulting model is better than the "parent" model, the original descriptor is discarded and the new descriptor is kept in its place. If the new model is not better, the original descriptor is kept. The best model after a defined number of iterations is then used to predict the test set activities.[48,49]

In this study, the RF and SVM algorithms available in R® 3.0.2 using the packages "e1071" and "randomForest" were implemented.[50] The SVM model was tuned to optimize performance. The *k*NN models were built using in-house modeling tools, also available at Chembench (chembench.mml.unc.edu).[51] Each method was performed with both MOE and Dragon descriptors in the QSAR model development process, as shown in the modeling workflow in **Figure 3**. The six resulting models were averaged to give a consensus prediction, as described in previous publications.[52,53] All models were validated using a five-fold cross validation. In this procedure, the training set was randomly split into five equal selected subsets. Four subsets (80%) were used as a training set and the compounds in the fifth subset (20%) were used as a test set. The training set was used to develop QSAR models and the resulting models were used to predict the test set. This procedure was repeated five times until all compounds were used in the test set once.[21,28]

Section 2: Results and Discussion

The modeling set was used to develop six individual QSAR models and their predictions were averaged as a consensus prediction. The model performance was indicated by five-fold cross validation of the modeling set itself and external prediction of a set of 264 unknown compounds. The performance was evaluated by calculating the sensitivity, specificity, and CCR for all models, as shown in **Figure 5**.

$$sensitivity = \frac{true\ positives}{(true\ positives + false\ negatives)}$$

$$specificity = \frac{true\ negatives}{(true\ negatives + false\ positives)}$$

$$CCR = \frac{sensitivity + specificity}{2}$$

These results were compared in order to evaluate the model performance for all individual models as well as the consensus model.

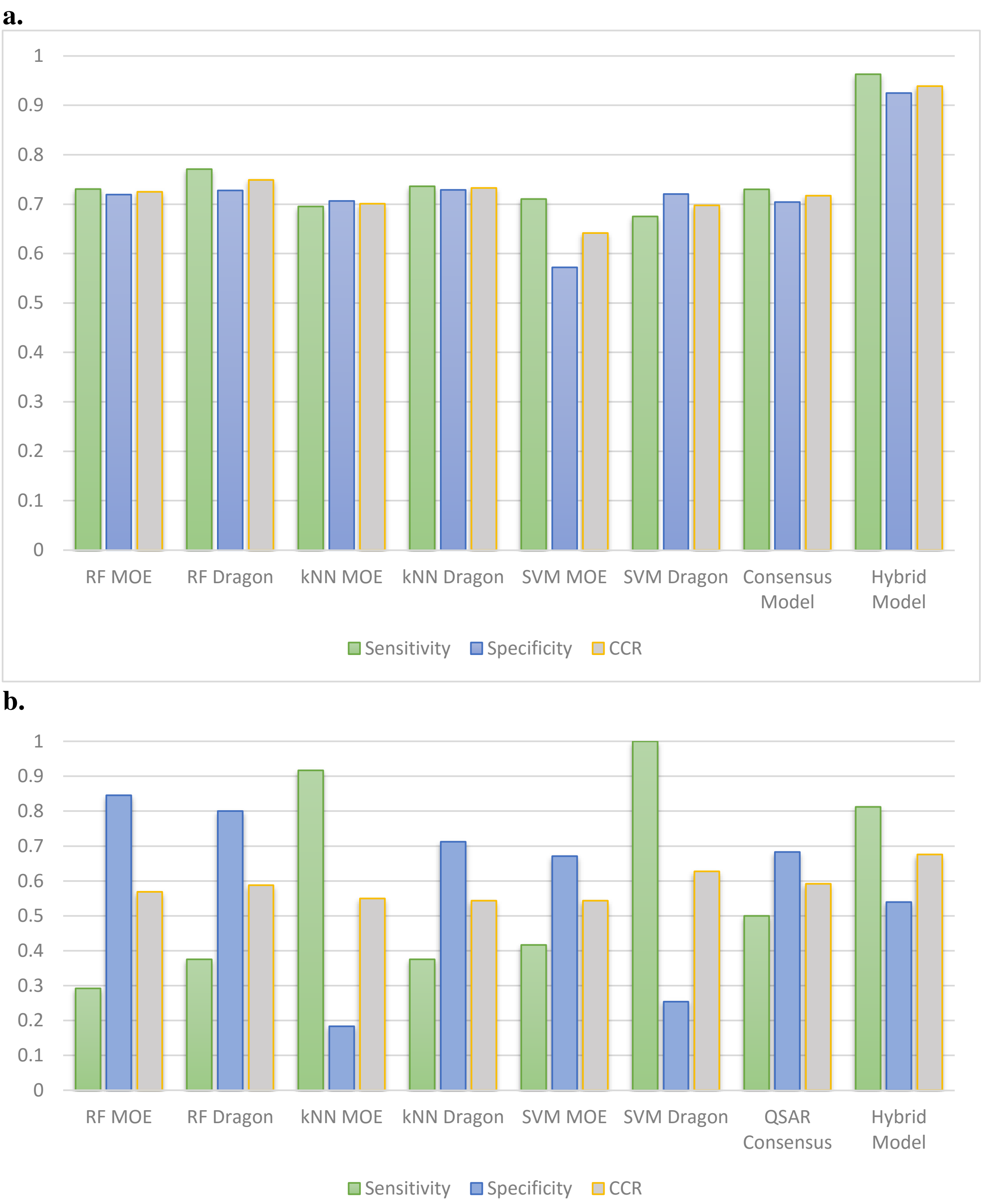


Figure 5: The performance of all resulting models: **a)** cross-validation of the 518 training set compounds; **b)** external validation of 264 unknown compounds.

Among the five-fold cross-validation procedures, the predictivity was similar across all the models (CCR = 0.642-0.749). However, the external predictions of the 264 unknown compounds showed a significant decrease in accuracy (CCR = 0.544-0.627), as observed in previous QSAR studies.[30,53] Among the individual cross-validation models, the Dragon RF model gave the best performance while the SVM MOE model returned the lowed CCR. The individual models' external predictions showed greater variety, with many of the models showing skewed performance toward sensitivity or specificity at the expense of the other. The kNN models had the lowest performance among the individual models, while the SVM Dragon model was shown to have the highest CCR (0.627).

Compared to individual models, the consensus model gave similar performance to the best individual models for both five-fold cross validation (sensitivity = 0.730, specificity = 0.704, and CCR = 0.717) and external predictions (sensitivity = 0.500, specificity = 0.683, and CCR = 0.592). In an attempt to improve the predictive ability of the QSAR model, an applicability domain was applied. An applicability domain consists of assessing the similarity of the target compound to its nearest neighbor. In this case, the Tanimoto coefficient was used to determine the similarity between compounds. When the distance between the compound and its nearest neighbor exceeds the set parameter, the activity of that target compound is not predicted.[30,54] Applying an applicability domain to both validation procedures did not show an improvement in predictive ability, so all predictions (100%) were retained when analyzing the QSAR models.

The cross-validation of the QSAR training set showed acceptable predictive ability, but the external prediction returned poor predictivity. In particular, some of the individual models showed unacceptably skewed results for the external prediction, with the SVM Dragon and *k*NN MOE giving high numbers of false positives (low specificity) and the Random Forest models giving high numbers of false negatives (low sensitivity). Although the consensus model shows relatively stable performance, its sensitivity of external test set prediction is much lower than cross validation due to the high proportion of false negatives. These results suggest the presence of activity cliffs within the data that may be addressed by the addition of biological response data to the computational model.

CHAPTER THREE: Biosimilarity Calculation and Hybrid Model Development

Section 1: Methods

Bioassay profiling similarity calculation

An in-house profiling tool[55] was used to extract relevant bioassay data from PubChem for each compound in the both the training and test sets. This tool identified each bioassay in PubChem for which the compound had an active, inactive, or inconclusive response and sorted the information into a matrix assigning values of 1, -1, or 0 for the possible responses. The two profiles were sorted and screened and all assays that were not present in both the training set and test set profiles were discarded. The PubChem assays in the profile were ranked by the numbers of active responses for the compounds in the training and test sets and filtered, as shown in the modeling workflow (**Figure 3**), to identify the most relevant bioassay profiles. The resulting PubChem bioassay profile consisted of 44 bioassays (shown in **Table 2**), which contain the largest number of active responses. The bioassay profile was screened to remove the original data set (AID 743077) so as to prevent redundancy in the similarity calculation. Bioassays were also removed if they did not contain useful experimental data or if they were redundant to another linked assay.

Table 2. PubChem bioassays used to create the bioprofile.

PubChem Bioassay	**Description**	**Active Responses**	**Inactive Responses**
AID 410	An assay for inhibitors and substrates of CYP1A2.	77	54
AID 686978	A screen to directly identify novel TDP1 inhibitors active in a cellular environment in the absence of CPT.	77	146
AID 1851	This assay detected inhibitors and substrates of various human cytochrome p450 (CYP450) isozymes	76	53

AID 504332	A qHTS assay for inhibitors of histone lysine methyltransferase G9a.	76	206
AID 686979	A screen to directly identify novel TDP1 inhibitors active in a cellular environment in the presence of CPT.	66	163
AID 884	An assay for inhibitors and substrates of CYP3A4.	60	131
AID 893	A qHTS assay for inhibitors of hydroxysteroid (17-beta) dehydrogenase 4 (HSD17B4)	35	188
AID 1030	A qHTS assay for inhibitiors of aldehyde dehydrogenase 1 (ALDH1A1)	35	186
AID 720532	A screen to identify small molecule inhibitors that block VSV-MARV (Marburg virus) binding or entry into cells	35	189
AID 899	An assay for inhibitors and substrates of CYP2C19.	34	88
AID 1490	A validation of the developed Sfp-PPtase assay that detects inhibition of Sfp-PPTase	33	210
AID 1883	An assay to identify inhibitors of malarial growth.	33	33
AID 504847	This HTS-compatible enzymatic assay that was developed to measure the inhibition between VDR and coregulator peptide SRC2-3 exerted by small molecules	33	311
AID 886	A qHTS assay for inhibitors of hydroxylacyl- coenzyme A dehydrogenase type II (HADH2)	29	163
AID 1460	An assay that screens for inhibitors of tau fibril formation	28	170
AID 891	An assay used for inhibitors and substrates of CYP2D6.	24	80
AID 894	A qHTS assay for inhibitiors of 15-hydroxyprostaglandin dehydrogenase (HPGD)	24	204
AID 883	An assay for inhibitors and substrates of CYP2C9.	23	97
AID 743244	A qHTS assay to identify gametocytocidal compounds which are capable of killing late stage P. falciparum gametocytes	23	211
AID 1996	This assay measures aqueous solubility of small molecule compounds.	22	7
AID 2147	A qHTS assay for inhibitors of human jumonji doma containing 2E (JMJD2E)	21	151
AID 2551	A qHTS assay to detect small molecule inhibitors of ROR gamma activity	20	245
AID 887	A qHTS assay for inhibitors of 15-human lipoxygenase	19	198

AID 588590	A high throughput replication assay to identify small molecule inhibitors of polymerase iota.	19	163
AID 720533	An assay to identify small molecule inhibitors that block VSV-LV binding or entry into cells	19	194
AID 589	A qHTS assay for spectroscopic profiling in the 4-MU spectral region	17	108
AID 590	A qHTS assay for spectroscopic profiling in the A-350 spectral region	17	107
AID 2549	A validation assay for inhibitors of Rec-Q-like DNA helicase 1 (RECQ1)	17	282
AID 504832	A qHTS for delayed death inhibitors of the malaria parasite plastid	16	130
AID 588579	A qHTS for inhibitors of DNA polymerase kappa	16	331
AID 588795	A qHTS for inhibitors of human flap endonuclease 1 (FEN1)	16	237
AID 1463	A counterscreen qHTS for inhibitors of tau fibril formation	15	188
AID 504339	A qHTS assay for inhibitors of JMJD2A-tudor domain	15	111
AID 596	A qHTS assay for tau filament binding	14	181
AID 485314	A qHTS for inhibitors of DNA polymerase beta	14	127
AID 902	A cytotoxicity assay to identify small molecular compounds that selectively target mutant p53-containing cancer cells	13	216
AID 880	A qHTS assay to identify inhibitors of the RGS:GPCR interaction	12	274
AID 915	A qHTS assay for the identification of small molecule antagonists for the hypoxia response element (HRE) signaling pathway	12	78
AID 2546	A counterscreen qHTS assay for inhibitors of ROR gamma transcriptional activity	12	242
AID 924	A cytotoxicity assay to identify small molecular compounds that selectively target mutant p53-containing cancer cells at the permissive temperature	11	221
AID 2472	Coupling assay counterscreen for a qHTS assay for inhibitors of fructose-1,6-bisphosphate aldolase from Giardia lamblia	11	186
AID 485317	HTS assay for inhibitors of augmenter of liver regeneration (ALR)	11	110
AID 1476	A qHTS assay for inhibitors of cruzain	10	226

AID 2517	A qHTS assay for inhibitors of human apurinic/apyrimidinic endonuclease (APE1)	10	161

The final profile was then used to calculate the biosimilarity between pairs of two compounds, as shown in the modeling workflow (**Figure 3**). This was done using the following WEBS formula:

$$Weighted\ Estimate\ of\ Biological\ Similarity\ (WEBS) = \frac{\sum(p + (\omega)n)}{\sum(p + \ (\omega)n + d)}$$

where *p* is the number of assays in which both compounds show active results, *n* is the number of assays in which both compounds show inactive results, and *d* is the number of assays in which the two compounds show opposite results. Inconclusive or missing data were not considered in the calculation. The negative response data (inactives) are weighted less than positive responses (actives) in the biosimilarity calculation. This is done because the positive (active) responses provide more useful information about biological similarity, since compounds may have inactive results for drastically different reasons. In this study, the weight parameter ω was given the value of 0.06. The resulting WEBS values range from 0-1 and were used to determine the nearest neighbors in the training set for each test set compound.

Each WEBS score was also assigned a confidence value which compared the number of assays included in the comparison with the total number of assays in the profile. This confidence value was used to prioritize compounds with matching WEBS scores when determining the biological nearest neighbors of a target compound. In determining the confidence value, as in the WEBS calculation, a shared active data point was given more weight than a shared inactive data point. Any compound with WEBS similarity score over

0.6 was considered as a potential nearest neighbor for the target compound. The ERα binding activities of up to the top five nearest neighbors were used to calculate the predicted activity of the relevant test set compound. The resulting predictions were rounded to assign a score of 0 (inactive) or 1 (active) to the test set compounds. When fewer than five nearest neighbors existed within the training set, all nearest neighbors were used. If there were no nearest neighbors (compounds with a WEBS score of at least 0.6) defined in the training set, no prediction was returned for that test set compound.

Hybrid Model Development

In order to form a hybrid model, the raw biosimilarity prediction was averaged with the QSAR consensus prediction for each compound, as shown in the modeling workflow (**Figure 3**).

$$\frac{Bioprediction + QSAR\ consensus}{2} = hybrid\ prediction$$

For compounds which were not able to be predicted by the biosimilarity tool due to missing data, the QSAR consensus prediction was used as the predicted value. Compounds with strongly diverging results from the QSAR consensus model and bioprediction were considered as inconclusive and removed. This method returned a prediction for 192 of the 264 test set compounds.

Section 2: Results

Bio-assay profile and predictions

Previous studies have shown improvements of QSAR models by incorporating biological data as extra descriptors into the modeling procedure. Relevant bioassay activity has been shown to be useful for improving the activity predictions when combined with QSAR methods.[52,56-58] Most recently, Wang et. al.[57] found that integrating specific biological activity descriptors into a hybrid model improved the predictive power. However, the lack of available bioassay data prevented this study from fully integrating the bioprofile from PubChem into the hybrid model. In this study, the in-house profiling tool was used to extract and optimize a biological profile containing 44 PubChem assays for 518 modeling set and 261 test set compounds. Using the WEBS score to calculate the biological similarity of each two compounds, those most similar compounds with WEBS scores over the nearest neighbor cut-off were identified for each test set compound and then used to predict the ERα binding potential. When combining the biosimilarity search with the QSAR consensus model as a hybrid model, the cross validation demonstrated a significant improvement of the accuracy over traditional QSAR modeling based only on chemical descriptors. Compared to the QSAR consensus model, the sensitivity, specificity and CCR of the hybrid model increased from 0.730 to 0.963, from 0.704 to 0.925, and from 0.717 to 0.939, respectively (**Figure 5**).

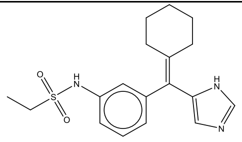
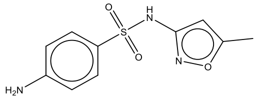
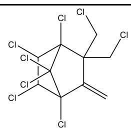
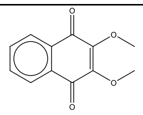
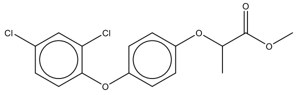
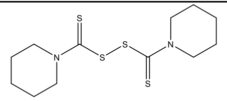
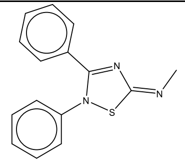
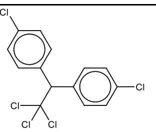
The external test set was also predicted using up to five of the most biosimilar compounds in the training set and the results combined with the consensus model to form a hybrid model. These predictions showed a noticeable improvement over the QSAR based solely on chemical descriptors. The external test set predictions returned a sensitivity = 0.813, specificity = 0.540, and CCR = 0.676 with a coverage of 73% (192 out of 264), an improvement over the QSAR prediction for which the consensus model had a sensitivity = 0.500, specificity = 0.683, and CCR = 0.592 (**Figure 5**). The increase of sensitivity in both cross validation and external predictions brings a considerable benefit when prioritizing potential EDCs for experimental testing.

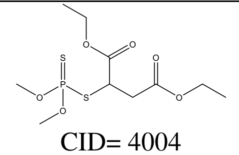
Section 3: Discussion

Considering the cross validation of the training set, the QSAR models all showed acceptable predictivity. However, the external prediction of 264 unknown compounds had significantly decreased prediction accuracy, especially for individual models. **Table 3** displays examples of compounds that were consistently incorrectly predicted by all QSAR models. The first active compound, N-[3-[cyclohexylidene(1H-imidazol-5-yl)methyl]phenyl]ethanesulfonamide (PubChem CID 6603710), an α-1D-adrenoceptor antagonist, is an ERα binder that was incorrectly predicted as inactive by all QSAR models. This compound's chemical nearest neighbor in the training set is the inactive compound sulfamethoxazole (PubChem CID 5329). Dimethoxynaphtoquinone (PubChem CID 3136) is also an active ERα binder that was incorrectly predicted by the QSAR consensus model. Its chemical nearest neighbor dichlofop-methyl (PubChem CID 39985) is an inactive compound in this assay. Similarly, the compound N-methyl-2,3-diphenyl-1,2,4-thiadiazol-5-imine (PubChem CID 682802) is an inactive compound. But its chemical nearest neighbor, dichlorodiphenyltrichloroethane (DDT) (PubChem CID 3036), in the training set is an ERα binder (active response). In each case, as expected, the QSAR model made a determination of activity based on compounds with similar chemical descriptors. However, for each of these examples, the most similar chemical compounds does not display the same biological activity. These prediction errors cannot be avoided if only chemical structure information is used for modeling.

The prediction of the test set improved when biosimilarity results were combined with the QSAR consensus model as a hybrid model. In particular, the sensitivity of the external test set prediction increased from 0.500 for the QSAR consensus model alone to 0.813 for the hybrid model. This indicates that the hybrid model was better able to correctly identify active compounds. The biological nearest neighbors, as determined by WEBS score, provide more useful information for the predictions of external compounds. For example, the biological nearest neighbor in the training set of N-[3-[cyclohexylidene(1H-imidazol-5-yl)methyl]phenyl]ethanesulfonamide (PubChem CID 6603710), an ERα binder, is toxaphene (PubChem CID 5284469), also an active compound (**Table 3**). The WEBS similarity score between these two compounds was 1.00. For the other external test set compounds in **Table 3**, their biological nearest neighbors show the same ERα binding activities as the relevant target compounds. Furthermore, the WEBS scores for these test set compounds show dissimilarity to their chemical nearest neighbors. For example, the inactive compound N-methyl-2,3-diphenyl-1,2,4-thiadiazol-5-imine (PubChem CID 682802) has a biological nearest neighbor (WEBS=1.00), malathion (PubChem CID 4004), a widely used insecticide that also showed inactive response in the ERα binding assay. Its chemical nearest neighbor, DDT (PubChem CID 3036), which is a now-banned insecticide, has a very low biosimilarity (WEBS =0.0169) to N-methyl-2,3-diphenyl-1,2,4-thiadiazol-5-imine. Seven PubChem assays, which have testing data for both compounds, show opposite results between these two compounds. The above analysis indicates that these pairs are activity cliffs, chemically similar compounds with different biological effects (in this case, ERα binding). However, the hybrid model, using biosimilarity as additional information in the modeling process, was able to correctly assign the appropriate activity to each test set compound.

Table 3: Three test set compounds (the first compound in each group) with their chemical nearest neighbor (the second compound) and biological nearest neighbor (the third compound).

	Compound	Activity	WEBS Score	Bioprofiles*
1	 CID= 6603710	Active	--	🟥🟥🟥🟦🟥🟥 *
	 CID= 5239	Inactive	0.117	🟦🟦🟦🟦🟦🟦
	 CID= 5284469	Active	1.00	⬜⬜🟥🟦🟥⬜
2	 CID= 3136	Active	--	🟥🟥🟥🟥🟥🟥 **
	 CID= 39985	Inactive	N/A	N/A
	 CID= 7188	Active	1.00	⬜⬜⬜🟥🟥🟥
3	 CID=682802	Inactive	--	🟥🟥🟥🟥🟥🟥🟥 ***
	 CID= 3036	Active	0.0169	🟦🟦🟦🟦🟦🟦🟦

	 CID= 4004	Inactive	1.00	🟥⬜⬜🟥🟥🟥🟥

* In the selected bioprofiles, the red color indicates active response, blue color indicates inactive response and white color indicates no data available.
The bioprofiles only consist of the assays out of 44 PubChem assays that have the data for the three compounds in each group:
First group bioprofile assays: PubChem AID 410, 883, 884, 893, 504832, 686978
Second group bioprofile assays: AID 410, 884, 504847, 686978, 686979, 743244
Third group bioprofile assays: AID 884, 886, 887, 893, 504847, 686978, 686979
N/A indicates there is no data available for this compound within these assays.

The bioassay response profile of the compounds shows promising potential to improve traditional QSAR models. Furthermore, when examining the PubChem assays used in the profile of this study, many targets of the assays regulate or are regulated by ERα. The highest ranked assay, which consists of the highest number of active responses for our training set compounds, was used to screen potential inhibitors of histone lysine methyltransferase G9a (PubChem AID 504332). This assay acts as a co-regulator in the estradiol-induced activation or repression of gene transcription by ERα.[59,60] Several other assays used in this profile specifically target enzymes in the cytochrome P450 (CYP450) family. The cytochrome P450 family consists of heme-containing oxidase enzymes, many of which are well-studied due to their involvement in the metabolism of many drugs. These assays include screening inhibitors for CYP1A2 (PubChem AID 410) and CYP3A4 (PubChem AID 884), and a composite screening results for various CYP450 inhibitors (PubChem AID 1851). These proteins modulate ERα signaling by helping to maintain the androgen/estrogen balance[61]. Through analyzing the bioassays within the response profile, it indicates the future direction of gathering useful data for evaluating potential ERα binders.

Section 4: Future Work and Conclusion

The biosimilarity methodology used in this project shows a promising way to improve the predictivity of traditional QSAR modeling. However, since many compounds may not have been tested and have no data available in public resources, the usefulness of biosimilarity is limited by its coverage. A potential strategy to address the limitation of missing data is by using "read-across" methods [62] to fill gaps in bioassay data for unknown compounds. While there are certain sources of uncertainty inherent in "read-across" methods, recent developments in "read-across" methods have moved toward the development of frameworks that allow for more certainty in gap filling. Another pitfall of using the public data is the prevalence of experimental errors and the redundancy between various assay results.[23,24] The model is only as good as the data used to develop it, and errors in database curation can lead to models with limited predictive utility. The next step in the development and refinement biosimilarity search tool will be to develop novel data mining approaches to which anticipate and address such issues.

The second current limitation lies in seamlessly incorporating chemical data with biological data. For many compounds, chemical descriptors are adequate for predicting activity. In fact, for a limited number of compounds, QSAR with chemical descriptors alone provided a better prediction than the hybrid model. In order to effectively incorporate bioassay profiling into QSAR predictions without severely limiting the applicable range of compounds, it is necessary to predict when chemical structure should be given more influence in the prediction and when that influence should be shifted to the biological profile. The integration of a biological profile alongside chemical descriptors is an important target for further study.

In this study, I first developed QSAR models for the qHTS assay data, which identify agonists for the ERα signaling pathway, provided in the Tox21 challenge. The external test set prediction of all QSAR models, including the consensus model, is lower than the cross validation results of the training set. However, by combining the biosimilarity search using the bioassay response profile automatically extracted from PubChem with the QSAR consensus predictions, a hybrid model was created. The resulting hybrid model showed a noticeable improvement in both cross-validation and external prediction results compared to QSAR models only based on chemical descriptors. This result demonstrated that integrating extra biological data in the modeling process can improve traditional QSAR models when predicting ERα binding potentials for unknown compounds. This strategy can be used to develop enhanced models to evaluate other types of toxicity for compounds of interest.

Supplemental Tables

Supplemental Table 1: The training set compounds identified by PubChem CID number along with their provided experimental activity, consensus prediction as determined by QSAR, prediction as determined by biosimilarity, and hybrid model prediction. Activity of 0 = inactive (nonbinder); activity of 1 = active (ERα binder); “--" indicates no prediction for the compound.

CID	Experimental Activity	QSAR Consensus Prediction	BioPrediction	Hybrid Prediction
30951	1	0.738	1	1
6702	1	0.546	1	1
2256	0	0.543	0	0
8425	0	0.779	1	1
6623	1	0.916	1	1
2347	1	0.767	1	1
7184	1	0.851	1	1
31404	0	0.112	0	0
16043	0	0.138	0	0
7768	0	0.453	0	0
6129	0	0.759	0	--
573	0	0.543	--	--
289	1	0.279	1	1
8371	1	0.186	1	--
15910	1	0.405	1	--
15331	0	0.508	0	0
4211	1	0.842	1	1
3036	1	0.644	1	1
76	1	0.925	--	1
540877	1	0.818	--	1
7070	1	0.541	1	1
15122	0	0.167	0	0
7069	0	0.264	0	0
949	0	0.524	0	0
4537	1	0.970	--	1
450	1	0.978	--	1
3285	1	0.977	--	1
12967	0	0.029	0	0
72060	0	0.453	0	0
3304	1	0.882	--	1
3468	1	0.451	--	--
522636	0	0.426	--	--
12089	0	0.108	0	0

3640	0	0.547	--	--
7638	1	0.841	1	1
8017	0	0.193	0	0
20240	1	0.839	1	1
5280863	1	0.909	1	1
4004	0	0.265	0	0
4080	1	0.967	--	1
13709	0	0.036	0	0
13622	1	0.210	1	--
9958	1	0.206	1	--
7543	0	0.844	0	--
21694	1	0.567	1	1
9074	1	0.758	1	1
7526	0	0.205	0	0
541197	1	0.887	--	1
6419	0	0.091	0	0
16741	0	0.525	0	0
8249	0	0.857	--	1
7103	1	0.713	1	1
4937	0	0.348	0	0
4947	1	0.542	1	1
5280343	1	0.760	1	1
5054	0	0.392	0	0
7276	0	0.317	0	0
5391	0	0.172	0	0
10544	0	0.460	--	--
8765	1	0.762	1	1
5284469	1	0.397	1	--
165628	1	0.878	--	1
3787925	1	0.960	--	1
10793	1	0.370	1	--
8256	1	0.649	1	1
13530	1	0.401	1	1
6619	1	0.581	1	1
335	0	0.172	0	0
7284	0	0.054	0	0
7321	0	0.056	0	0
7504	0	0.278	0	0
8609	1	0.736	1	1
8914	0	0.790	0	--
61163	0	0.122	0	0
13254	1	0.723	1	1
5354198	1	0.825	1	1

24874	0	0.190	0	0
5280961	1	0.923	1	1
5281708	1	0.933	1	1
5280378	1	0.909	1	1
22283	1	0.690	--	1
6025	1	0.512	1	1
5408	1	0.890	--	1
13089	1	0.468	1	1
8814	1	0.481	1	1
15	1	0.947	--	1
3269	1	0.931	--	1
698	1	0.973	--	1
3606	1	0.935	--	1
3049	1	0.944	--	1
5280443	1	0.928	1	1
4788	1	0.892	1	1
5280373	1	0.924	1	1
5281607	1	0.928	0	--
4632	1	0.746	1	1
8572	1	0.822	1	1
6626	1	0.683	1	1
69150	1	0.911	0	--
66166	1	0.911	1	1
12111	1	0.931	1	1
66030	1	0.530	1	1
3405	0	0.798	--	1
14138	1	0.858	1	1
107377	1	0.641	1	1
7180	1	0.749	1	1
7175	1	0.543	1	1
11742	1	0.827	1	1
2040	1	0.805	1	1
2096	1	0.889	--	1
3698	0	0.452	0	0
223407	1	0.939	--	1
2524	0	0.245	--	0
10580	1	0.403	--	0
430461	1	0.848	0	0
2769	0	0.535	0	0
16351	0	0.548	0	0
2913	0	0.059	0	0
2949	1	0.871	--	1
4659180	1	0.926	--	1

3159	1	0.533	--	--
3262	1	0.942	--	1
3263	1	0.961	--	1
3265	1	0.871	--	1
3590886	1	0.967	--	1
3267	1	0.972	--	1
699	1	0.212	--	0
3288	1	0.898	--	1
3415	1	0.299	--	0
628035	1	0.920	--	1
3647	0	0.216	0	0
3728	0	0.066	0	0
3735	1	0.300	1	1
3961	0	0.626	0	0
522463	1	0.899	--	1
4197	0	0.840	0	--
4240	0	0.697	--	1
220503	1	0.915	--	1
4432	1	0.929	--	1
4435	1	0.936	--	1
4536	1	0.979	--	1
4567	0	0.231	0	0
5407	1	0.785	--	1
5434	0	0.466	--	--
5470	1	0.898	--	1
5472	1	0.796	1	1
5479	1	0.792	1	1
4659569	0	0.459	0	0
5611	0	0.442	--	--
4638	1	0.781	--	1
36324	0	0.096	0	0
8413	1	0.484	1	1
8076	0	0.161	0	0
37517	0	0.370	0	0
31368	0	0.275	0	0
29732	0	0.764	1	1
12901	0	0.470	0	0
26124	0	0.791	1	1
5543	0	0.173	0	0
10907	0	0.315	0	0
6846605	0	0.155	--	0
2202	0	0.771	0	--
542762	0	0.270	0	0

7236	0	0.687	0	0
7374	0	0.124	0	0
73773	1	0.833	1	1
7258	1	0.112	1	--
7411	0	0.127	0	0
6641	1	0.614	1	1
6777	0	0.058	0	0
409778	1	0.420	1	1
7993	0	0.348	0	0
6589	0	0.486	0	0
11764	0	0.100	1	--
66666	0	0.149	0	0
3220	1	0.658	1	1
12994	1	0.459	--	--
22463	1	0.830	1	1
7546	1	0.892	1	1
8430	0	0.038	0	0
25711	0	0.247	0	0
13195	0	0.097	0	0
16097	0	0.041	0	0
6753	1	0.150	1	--
11876	0	0.262	0	0
6829	0	0.159	0	0
3543259	0	0.127	0	0
4761	0	0.453	--	--
7013	1	0.647	1	1
10129	0	0.164	0	0
1057	1	0.286	--	0
1070	1	0.352	--	0
5297	0	0.145	--	0
5329	0	0.423	0	0
6618	1	0.378	1	--
6895	0	0.532	0	0
18725	0	0.202	0	0
111037	0	0.790	0	--
6934	0	0.086	0	0
736	0	0.104	--	0
7104	0	0.190	1	--
7355	0	0.063	0	0
7412	0	0.543	0	0
7495	1	0.209	1	--
8663	1	0.319	1	1
68191	0	0.051	0	0

11369	0	0.133	0	0
12388	0	0.174	0	0
12573	0	0.291	0	0
70400	1	0.786	1	1
19165	0	0.314	0	0
78501	0	0.620	0	0
21984	0	0.056	0	0
22206	1	0.162	1	--
64819	1	0.460	1	1
23284	1	0.311	0	0
93079	0	0.351	0	0
26295	1	0.735	1	1
27423	1	0.522	1	1
28777	0	0.152	0	0
87323	1	0.188	1	--
62530	1	0.213	1	--
171144	1	0.352	1	1
91604	0	0.162	0	0
64865	0	0.178	0	0
4128060	1	0.334	1	1
338733	1	0.384	1	1
7112	1	0.918	1	1
7344	0	0.401	0	0
8858	0	0.086	0	0
75103	0	0.103	0	0
74000387	0	0.651	--	1
86583	0	0.268	0	0
22833454	0	0.170	0	0
320760	0	0.451	--	--
13551	0	0.086	0	0
12251	0	0.047	0	0
5056	1	0.883	--	1
6527	0	0.680	1	1
3031	0	0.163	0	0
123626	0	0.053	--	0
43226	1	0.892	1	1
86398	1	0.788	1	1
31099	0	0.166	--	0
92421	0	0.052	0	0
213031	1	0.862	1	1
91731	1	0.764	1	1
39985	0	0.153	0	0
11432	0	0.454	0	0

2160	0	0.670	0	0
2683	0	0.432	--	0
2958	1	0.767	--	1
12324	0	0.050	--	0
91649	0	0.740	1	1
4160	1	0.870	--	1
3408	1	0.707	--	1
4092426	1	0.949	--	1
15546	0	0.308	0	0
29025	0	0.542	--	--
4542	1	0.922	--	1
439222	1	0.665	--	1
5410	1	0.914	--	1
1027	1	0.764	1	1
2681	0	0.782	--	1
25146	1	0.829	1	1
90571	1	0.633	1	1
73864	1	0.653	1	1
70837	1	0.755	1	1
31208	1	0.893	1	1
21804	1	0.744	1	1
8208	0	0.847	0	--
92387	1	0.854	1	1
8570	1	0.805	1	1
69785	1	0.553	1	1
11442	0	0.388	0	0
8861	0	0.310	0	0
3752	0	0.485	--	--
13783	1	0.284	1	1
32490	0	0.700	0	0
5513	1	0.302	1	1
4745	0	0.438	--	--
3686	0	0.216	1	--
3430	0	0.544	--	--
9934458	0	0.633	0	0
3054	1	0.959	1	1
86705	1	0.867	--	1
3454	0	0.260	0	0
4456	0	0.762	--	1
14896	0	0.148	0	0
107080	1	0.790	1	1
8571	1	0.884	1	1
1066	0	0.229	0	0

15325	1	0.516	1	1
83268	0	0.032	0	0
77328	1	0.353	1	1
13345	0	0.239	0	0
100524	1	0.558	1	1
40520	1	0.893	1	1
3435	1	0.706	--	1
80900	0	0.614	0	0
91683	1	0.889	1	1
62556	1	0.575	1	1
17435	0	0.276	0	0
4090	0	0.301	--	0
3341	1	0.895	1	1
8096	0	0.189	0	0
2678	0	0.539	--	--
7738	0	0.616	0	0
96359	0	0.070	0	0
69311	0	0.663	0	0
14127	1	0.683	1	1
67005	1	0.281	1	--
25622	1	0.137	1	--
3482402	0	0.190	0	0
86607	1	0.857	1	1
75282	1	0.231	1	--
6785831	0	0.064	--	0
92667	0	0.386	0	0
7388	0	0.491	--	--
22628	0	0.364	0	0
8420	0	0.680	0	0
103005	1	0.639	1	1
243274	1	0.844	1	1
73852	1	0.860	1	1
75557	0	0.128	0	0
75576	1	0.887	1	1
117640	1	0.748	1	1
170286	1	0.764	1	1
93312	0	0.133	0	0
1923	1	0.935	1	1
409069	1	0.879	--	1
6729	0	0.101	--	0
3158	0	0.874	--	1
3676	0	0.628	0	0
5154	1	0.798	1	1

2310	0	0.594	0	0
4641	0	0.807	1	1
5156	0	0.495	--	--
3243	1	0.655	--	1
5531	0	0.348	0	0
3326	0	0.425	0	0
4912	0	0.220	0	0
4857	0	0.304	--	0
73342	0	0.433	--	--
2356	0	0.777	--	1
1302	0	0.197	--	0
3066	0	0.115	--	0
2120	0	0.303	0	0
65854	0	0.126	0	0
5278	0	0.070	0	0
33925	0	0.690	0	0
71350	0	0.556	--	--
424631	0	0.373	--	0
5244	0	0.325	--	--
3344	0	0.085	--	--
3780	0	0.267	--	--
4868	0	0.145	--	--
3362	0	0.703	0	0
3748	1	0.724	--	1
13675140	1	0.973	--	1
2158	0	0.455	--	--
56208	0	0.292	--	0
666418	0	0.130	0	0
6293	1	0.600	1	1
2806	1	0.804	--	--
4432690	1	0.182	--	0
61574	0	0.115	0	0
2474	0	0.142	--	0
31477	1	0.618	1	1
120081	0	0.419	0	0
41287	0	0.288	--	0
32593	0	0.794	0	--
28718	0	0.648	0	0
10832	0	0.205	0	0
65630	0	0.607	0	0
517915	1	0.887	--	1
4398807	1	0.822	--	1
4901	1	0.770	--	--

67686	0	0.815	1	1
1236	0	0.177	0	0
262961	1	0.773	--	1
10275	0	0.353	0	0
71874	0	0.042	--	0
28554	0	0.798	0	--
68770	0	0.645	0	0
68733	1	0.820	--	1
60867	0	0.577	--	--
4006	0	0.147	0	0
2999	1	0.982	--	1
65947	1	0.792	1	1
3033226	0	0.559	0	0
23674	1	0.532	1	1
863	1	0.682	1	1
5771	1	0.854	--	1
1451	1	0.894	--	1
4971	1	0.875	--	1
3977	1	0.180	--	0
8230	1	0.763	--	1
60651	0	0.446	0	0
4619	1	0.917	--	1
540766	1	0.893	--	1
2624	0	0.111	--	0
53394893	1	0.256	--	0
4237	0	0.836	1	1
4089	1	0.772	--	1
14883207	1	0.890	--	1
11622909	0	0.441	--	--
53461729	1	0.878	--	1
85823	1	0.918	--	1
3148	0	0.363	--	0
76513276	0	0.172	--	0
248862	0	0.242	--	0
457814	0	0.238	--	0
2759	0	0.324	--	0
266651	0	0.565	--	--
13916	1	0.273	1	--
13118	1	0.753	1	1
3416	1	0.780	--	1
6426711	0	0.496	--	--
89386	1	0.878	--	1
11723708	1	0.856	1	1

27211	1	0.820	1	1
75346	0	0.231	--	0
9025	0	0.293	0	0
17166	0	0.078	--	0
70917	0	0.163	0	0
521196	0	0.378	0	0
8542	1	0.846	--	1
13132266	0	0.176	0	0
6873	0	0.582	0	0
8834	1	0.021	1	--
31353	1	0.542	1	1
31954	0	0.229	0	0
44568380	0	0.091	--	0
9885850	0	0.883	1	1
22661774	0	0.326	--	0
10233356	1	0.630	1	1
53931237	1	0.593	--	1
72679334	0	0.147	--	0
9909677	1	0.907	1	1
53316387	0	0.090	--	0
3247	1	0.951	--	1
96088	1	0.528	1	1
42504	0	0.808	0	--
75547	1	0.535	1	1
61384	0	0.057	0	0
61950	0	0.068	--	0
10229	1	0.206	1	--
833466	0	0.870	0	--
7921	0	0.083	0	0
18522	0	0.388	0	0
3685	1	0.763	1	1
411697	0	0.597	--	1
13505	1	0.309	0	0
83050	1	0.659	1	1
87250	1	0.760	1	1
586708	1	0.811	1	1
56638112	1	0.827	1	1
23019	1	0.917	1	1
608116	1	0.946	1	1
8633	1	0.643	1	1
919792	1	0.823	1	1
14642	1	0.781	1	1
17570	1	0.916	1	1

123504	1	0.917	1	1
7563	1	0.821	1	1
232446	1	0.924	1	1
79717	1	0.507	1	1
12472902	1	0.835	1	1
5993	1	0.270	1	1
190373	0	0.118	--	0
9824345	0	0.080	--	0
516981	0	0.020	--	0
69144	0	0.444	0	0
82703	1	0.267	1	--
9884915	1	0.590	1	1
3425	1	0.417	--	--
12940545	0	0.143	--	0
4641498	1	0.845	--	1
3033538	1	0.451	--	--
32603	0	0.345	--	--
562114	1	0.899	--	--
7188	1	0.540	1	1
65632	0	0.072	0	0
2227	1	0.769	--	1
53315588	1	0.523	--	--
11338777	0	0.481	--	--
76007663	0	0.074	--	0
1999	0	0.353	0	0
6179	0	0.363	--	0
4921	1	0.671	1	1
13324	1	0.950	--	1
107786	0	0.720	1	1
19527	0	0.093	--	0
31331	0	0.152	1	--
15129	0	0.196	--	0
62380	0	0.102	--	0
15686	1	0.711	--	1
156414	1	0.832	1	1
10127622	1	0.845	0	--
3268	1	0.595	--	1
975	1	0.521	--	--
4079	1	0.673	--	1
3266	1	0.971	--	1
65388	0	0.084	--	0
19961652	1	0.828	--	1
3510	0	0.064	--	0

4893	0	0.612	0	0
8446	0	0.054	0	0
26204	0	0.858	0	--
27435	0	0.191	--	0
9164	0	0.503	1	1
3510569	0	0.282	--	0
3014138	0	0.047	0	0
387179	0	0.839	1	1
11796	1	0.597	1	1
2737408	1	0.324	1	1
71593	0	0.381	0	0
44373822	0	0.982	0	--

Supplemental Table 2: The test set compounds identified by PubChem CID number along with their provided experimental activity, consensus prediction as determined by QSAR, prediction as determined by biosimilarity, and hybrid model prediction. Activity of 0 = inactive (nonbinder); activity of 1 = active (ERα binder); “--" indicates no prediction for the compound.

CID	Experimental Activity	QSAR Consensus Prediction	BioPrediction	Hybrid Prediction
1868	0	0.402	0	0
2020	0	0.364	0.25	0
241893	0	0.451	0.5	--
6604918	0	0.529	0.5	--
1793	0	0.361	0.75	1
9572720	0	0.522	--	--
5131	0	0.457	--	--
4498	0	0.210	0.5	0
903	0	0.523	0.5	--
1795	1	0.412	0.75	1
2259	0	0.601	1	1
1367	0	0.305	0.25	0
6603710	1	0.411	0.75	1
1150	0	0.340	0	0
1720	0	0.430	0.33333333	0
2207	0	0.272	0.75	--
3857	0	0.307	0.75	--
173615	0	0.418	--	0
1725	0	0.328	0.25	0
1820	0	0.421	0.75	1
2039	0	0.257	1	1
1564	0	0.405	--	0
97587	0	0.244	0.25	0
1365	0	0.370	0.25	0
469	0	0.252	--	0
4615193	0	0.414	0.75	1
1579	0	0.455	--	--
10061214	0	0.628	0.5	1
70547	0	0.379	0.25	0
11401613	0	0.622	0.5	1
2071	0	0.384	0.75	1
4043357	0	0.373	--	0
1961	0	0.682	--	1
1232	0	0.215	--	0

1908	0	0.386	0.75	1
2252	0	0.438	--	0
5149739	0	0.533	--	--
73153239	0	0.611	--	1
6603998	0	0.358	0.5	0
15186066	0	0.323	--	0
6603717	0	0.585	0.25	0
56965900	1	0.587	--	1
5372720	0	0.392	0.25	0
73153241	0	0.504	--	--
5024764	1	0.487	0.5	--
73153243	0	0.467	--	--
1248	0	0.444	--	0
1858	0	0.411	0.5	--
3519541	0	0.493	0.25	0
2302	1	0.514	--	--
16219010	0	0.499	--	--
89105	0	0.444	0.75	1
2491	0	0.502	--	--
2419	0	0.408	0.25	0
4217	0	0.452	0.75	1
108107	0	0.401	0.75	1
73153244	0	0.488	--	--
5016	0	0.603	--	1
21157	0	0.333	0.5	0
2921148	0	0.418	0.5	--
53421694	0	0.354	--	0
4059895	0	0.518	0.75	1
2703	1	0.569	1	1
5097	0	0.411	--	0
4545575	0	0.385	--	0
108042	0	0.701	--	1
119376	0	0.379	0.25	0
291704	1	0.469	--	--
8743	0	0.350	0.5	0
5126051	0	0.468	--	--
26532	0	0.386	--	0
1959	0	0.504	--	--
73153248	0	0.419	--	0
5280569	0	0.539	1	1
2692	0	0.374	0.75	1
5685	0	0.388	--	0
1551	0	0.256	--	0

11957508	0	0.443	--	0
1774	0	0.243	0.25	0
1548	1	0.465	--	--
3068	1	0.515	--	--
1335	0	0.215	--	0
11957516	0	0.574	0.5	--
1271	0	0.278	--	0
1329	0	0.406	0	0
501	0	0.280	--	0
4687	0	0.273	0.5	0
1225	0	0.609	0.75	1
1902	0	0.430	--	0
4278	0	0.284	0.5	0
3136	1	0.411	0.75	1
4712	0	0.612	0.75	1
5353574	0	0.418	0.75	1
1734	0	0.364	0.5	0
547	0	0.446	0.5	--
3153	0	0.465	0.75	1
161713	0	0.346	0.5	0
73153253	0	0.530	--	--
161930	0	0.521	0.5	--
30137	0	0.645	--	1
1242	0	0.643	0.5	1
108137	0	0.537	0.25	0
1609	0	0.451	0.25	0
6603792	0	0.515	0.75	1
6412645	0	0.383	0.75	1
132496	0	0.538	0.75	1
5019	0	0.683	0.75	1
1317	0	0.249	0.25	0
4474781	0	0.769	0	0
6603827	0	0.459	0.25	0
275	0	0.317	0.5	0
6603857	0	0.498	0.5	--
5034	0	0.625	1	1
11623092	0	0.576	--	1
612745	0	0.406	--	0
3213	1	0.648	1	1
4375	0	0.441	0.25	0
6603849	1	0.423	0.5	--
2105	0	0.308	--	0
1126109	0	0.444	0.5	--

9819328	0	0.588	--	1
3423	0	0.561	0.75	1
3456	0	0.473	1	1
5425	0	0.264	0.5	0
1742	0	0.391	0.25	0
3532	0	0.229	0.5	0
107812	0	0.215	0.75	--
6422124	0	0.491	--	--
3585	0	0.424	0.75	1
6603882	0	0.387	0.75	1
1221	0	0.384	0.75	1
11957563	0	0.409	--	0
4266	0	0.321	0.5	0
94945	0	0.417	--	0
1245	0	0.481	--	--
1745	0	0.229	0.75	--
5202	0	0.503	0.5	--
2817242	0	0.633	0.75	1
9572720	0	0.500	--	--
1219	0	0.535	--	--
3615	0	0.209	--	0
128018	0	0.478	0.75	1
4529080	0	0.527	--	--
4182	0	0.402	--	0
262093	1	0.502	1	1
6603884	0	0.538	0.33333333	0
3538	0	0.573	0.5	--
1738	0	0.477	0.75	1
107794	0	0.490	0.75	1
54722180	0	0.697	--	1
1220	0	0.537	0.5	--
4216	0	0.461	0.75	1
3802	0	0.396	0.5	0
73153258	1	0.540	--	--
1352	0	0.403	--	0
824226	0	0.593	0.75	1
6272	0	0.485	--	--
3667	0	0.547	--	--
3668	1	0.393	--	0
425	1	0.791	--	1
1719	0	0.391	--	0
4203080	0	0.408	--	0
14973220	0	0.365	--	0

10729	0	0.370	0	0
3845	0	0.490	0.25	0
3885	0	0.556	1	1
9907994	0	0.407	--	0
5281672	0	0.728	1	1
133633	0	0.513	0.75	1
1222	0	0.381	0.75	1
9837540	0	0.597	--	1
3614	0	0.217	0.75	--
1869	0	0.362	--	0
74339046	0	0.451	--	--
123679	0	0.429	0.75	1
155806	0	0.351	0	0
6603928	0	0.295	0.75	--
10035933	0	0.420	0.5	--
24906282	1	0.601	--	1
17756950	1	0.599	0.66666667	1
4376	0	0.225	0.5	0
6843761	0	0.504	--	--
16759251	0	0.390	0.25	0
107982	0	0.525	0.5	--
3661570	0	0.608	0.75	1
5311200	0	0.477	0.75	1
16118	0	0.294	--	0
6603828	0	0.427	0.25	0
4023	0	0.546	--	--
4533	0	0.610	--	1
2856102	0	0.587	--	1
3514	0	0.423	--	0
4420320	0	0.441	--	0
93004	0	0.304	0.25	0
1390	0	0.203	0	0
11957638	0	0.487	--	--
5129	0	0.455	--	--
1893	0	0.377	0.25	0
2378095	0	0.629	--	1
5079496	0	0.227	--	0
1237	0	0.435	--	0
4108	0	0.470	1	1
1894361	0	0.540	0.75	1
5353329	0	0.456	0.75	1
132787	0	0.465	0.75	1
4431	0	0.431	0.75	1

4813	1	0.783	--	1
3644637	0	0.487	--	--
126402	0	0.449	0.5	--
6811971	0	0.515	--	--
24906273	0	0.460	--	--
1641	0	0.413	0.5	--
1585	1	0.386	--	0
5309	0	0.400	--	0
23643664	0	0.557	0.75	1
4814	0	0.403	0.25	0
3674	1	0.460	--	--
2301	1	0.435	--	0
71367729	1	0.727	--	1
1960	0	0.431	0.75	1
5117	0	0.533	--	--
1209	0	0.337	--	0
9799515	0	0.455	--	--
3495594	0	0.543	1	1
2779853	1	0.398	1	1
11983295	0	0.455	--	--
123981	0	0.362	0.25	0
443751	0	0.538	--	--
692413	0	0.526	--	--
--	0	0.490	--	--
313280	0	0.662	--	1
1103	0	0.307	0.5	0
682802	0	0.509	0.25	0
4205032	0	0.451	--	--
4564402	0	0.421	0.5	--
10220536	0	0.656	0.75	1
5282	0	0.319	0.75	--
4389	0	0.345	--	0
3135	0	0.463	--	--
10336538	0	0.486	--	--
2055	0	0.645	--	1
92409	0	0.454	--	--
6849066	0	0.635	--	1
1218	0	0.661	0.75	1
44828493	0	0.443	0.75	1
327653	0	0.364	1	1
14032955	0	0.374	--	0
2056	0	0.589	--	1
44287897	0	0.466	--	--

73153275	0	0.696	--	1
398148	0	0.366	0.75	1
5637	0	0.554	--	1
5714	0	0.444	0.75	1
4518925	0	0.630	--	1
5631	0	0.634	--	1
3862	0	0.370	0.5	0
2940	0	0.494	--	--
15897	0	0.412	--	0
3725	0	0.676	--	1
62824	0	0.319	1	1
2057	0	0.727	--	1
5619	0	0.438	--	0
521106	0	0.763	0.75	1
5640	0	0.395	0.25	0

References

1. Hall, J. M.; Couse, J. F.; Korach, K. S. The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J. Biol. Chem.* **2001***, 276*, 36869-36872.

2. Mangelsdorf, D. J.; Thummel, C.; Beato, M.; Herrlich, P.; Schütz, G.; Umesono, K.; Blumberg, B.; Kastner, P.; Mark, M.; Chambon, P.; Evans, R. M. The nuclear receptor superfamily: The second decade. *Cell* **1995***, 83*, 835-839.

3. Kuiper, G. G. J. M.; Enmark, E.; Gustafsson, J. Å.; Nilsson, S.; Pelto-Huikko, M. Cloning of a novel estrogen receptor expressed in rat prostate and ovary. *Proc. Natl. Acad. Sci. U. S. A.* **1996***, 93*, 5925-5930.

4. Hewitt, S. C.; Harrell, J. C.; Korach, K. S. Lessons in estrogen biology from knockout and transgenic animals. *Annu. Rev. Physiol.* **2005***, 67*, 285-308.

5. Glass, C. K. Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers. *Endocr. Rev.* **1994***, 15*, 391-407.

6. Shanle, E. K.; Xu, W. Endocrine disrupting chemicals targeting estrogen receptor signaling: Identification and mechanisms of action. *Chem. Res. Toxicol.* **2011***, 24*, 6-19.

7. Schug, T. T.; Janesick, A.; Blumberg, B.; Heindel, J. J. Endocrine disrupting chemicals and disease susceptibility. *J. Steroid Biochem. Mol. Biol.* **2011***, 127*, 204-215.

8. Coleman, K. M.; Smith, C. L. Intracellular signaling pathways: Nongenomic actions of estrogens and ligand-independent activation of estrogen receptors. *Frontiers in Bioscience- Landmark* **2001***, 6*, D1379-D1391.

9. Safe, S. Transcriptional activation of genes by 17 beta-estradiol through estrogen receptor-Sp1 interactions. *Vitam. Horm.* **2001***, 62*, 231-252.

10. Loven, M. A.; Wood, J. R.; Nardulli, A. M. Interaction of estrogen receptors α and β with estrogen response elements. *Mol. Cell. Endocrinol.* **2001***, 181*, 151-163.

11. Klopman, G.; Chakravarti, S. K. Structure–activity relationship study of a diverse set of estrogen receptor ligands (I) using MultiCASE expert system. *Chemosphere* **2003***, 51*, 445-459.

12. Klopman, G.; Chakravarti, S. K. Screening of high production volume chemicals for estrogen receptor binding activity (II) by the MultiCASE expert system. *Chemosphere* **2003***, 51*, 461-468.

13. Swaby, R. F.; Sharma, C.; Jordan, V. C. SERMs for the treatment and prevention of breast cancer. *Reviews in Endocrine & Metabolic Disorders* **2007***, 8*, 229-239.

14. Koehler, K. F.; Helguero, L. A.; Haldosén, L.; Warner, M.; Gustafsson, J. Reflections on the discovery and significance of estrogen receptor beta. *Endocr. Rev.* **2005***, 26*, 465-478.

15. Mueller, S. O.; Korach, K. S. Estrogen receptors and endocrine diseases: lessons from estrogen receptor knockout mice. *Curr Opin Pharmacol* **2001***, 1*, 613-619.

16. Committee on Toxicity Testing and Assessment of Environmental Agents, N.R.C. *Toxicity testing in the 21st century: A vision and a strategy;* The National Academies Press: Washington, D.C., 2007.

17. Zhu, H.; Zhang, J.; Kim, M. T.; Boison, A.; Sedykh, A.; Moran, K. Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol.* **2014***, 27*, 1643-1651.

18. Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative High-Throughput Screening: A Titration-Based Approach That Efficiently Identifies Biological Activities in Large Chemical Libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, 11473-11478.

19. US Food and Drug Administration Endocrine Disruptor Knowledge Database (EDKB). http://www.fda.gov/ScienceResearch/BioinformaticsTools/EndocrineDisruptorKnowledgebase/default.htm (accessed 2015).

20. Valerio, L. G. Review: In silico toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol.* **2009***, 241*, 356-370.

21. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* **2003***, 17*, 241-253.

22. Stouch, T. R.; Johnson, S. R.; Doweyko, A.; Li, Y.; Chen, X. -.; Kenyon, J. R. In silico ADME/Tox: Why models fail. *J. Comput. Aided Mol. Des.* **2003***, 17*, 83-92.

23. Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010***, 50*, 1189-1204.

24. Young, D.; Martin, T.; Harten, P.; Venkatapathy, R. Are the chemical structures in your QSAR correct? *QSAR and Combinatorial Science* **2008***, 27*, 1337-1345.

25. Klopman, G.; Zhu, H.; Fuller, M. A.; Saiakhov, R. D. Searching for an enhanced predictive tool for mutagenicity. *SAR QSAR Environ. Res.* **2004***, 15*, 251-263.

26. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **1901**, *2*, 559-572.

27. Maggiora, G. M. On outliers and activity cliffs--why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

28. Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494-3504.

29. Richard, A. M. Future of toxicology--predictive toxicology: An expanded view of "chemical toxicity". *Chem. Res. Toxicol.* **2006**, *19*, 1257-1262.

30. Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatical, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766-784.

31. Zhang, L.; Zhu, H.; Golbraikh, A.; Tropsha, A.; Oprea, T. I. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* **2008**, *25*, 1902-1914.

32. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.* **2005**, *33*, 445-459.

33. Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. Navigating structure-activity landscapes. *Drug Discov. Today* **2009**, *14*, 698-705.

34. Schultz, T. W.; Cronin, M. T. D.; Walker, J. D.; Aptula, A. O. Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *J.Mol. Struct.* **2003**, *622*, 1-22.

35. Benigni, R.; Bossa, C. Predictivity and Reliability of QSAR Models: The Case of Mutagens and Carcinogens. *Toxicology Mechanisms & Methods* **2008**, *18*, 137-147.

36. Liu, H.; Papa, E.; Gramatica, P. Evaluation and QSAR modeling on multiple endpoints of estrogen activity based on different bioassays. *Chemosphere* **2008**, *70*, 1889-1897.

37. Taha, M. O.; Tarairah, M.; Zalloum, H.; Abu-Sheikha, G. Pharmacophore and QSAR modeling of estrogen receptor β ligands and subsequent validation and in silico search for new hits. *J. Mol. Graph. Model.* **2010**, *28*, 383-400.

38. Zhang, L.; Sedykh, A.; Tripathi, A.; Zhu, H.; Afantitis, A.; Mouchlis, V. D.; Melagraki, G.; Rusyn, I.; Tropsha, A. Identification of putative estrogen receptor-mediated endocrine disrupting chemicals using QSAR- and structure-based virtual screening approaches. *Toxicol. Appl. Pharmacol.* **2013***, 272*, 67-76.

39. Deng, C. L.; Chen, X. X.; Lu, H. Y.; Yang, X.; Luan, F.; Cordeiro, M. Prediction of the Estrogen Receptor Binding Affinity for both hER(alpha) and hER(beta) by QSAR Approaches. *Lett. Drug Des. Disc.* **2014***, 11*, 265-278.

40. Zang, Q.; Rotroff, D. M.; Judson, R. S. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. *J. Chem. Inf. Model.* **2013***, 53*, 3244-3261.

41. Scior, T.; Medina-Franco, J.; Do, Q. T.; Martinez-Mayorga, K.; Rojas, J.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009***, 16*, 4297-4313.

42. Johnson, S. R. The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **2008***, 48*, 25-26.

43. Cruz-Monteagudo, M.; Medina-Franco, J.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N.; Borges, F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **2014***, 19*, 1069-1080.

44. PubChem BioAssay: AID 743077. https://pubchem.cnbi.nlm.nih.gov/assay/assay.cgi?aid=7430772015).

45. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006***, 11*, 1046-1053.

46. Vapnik, V. *The nature of statistical learning theory.* Springer Science & Business Media: 2000.

47. Breiman, L. Random forests. *Mach. Learning* **2001***, 45*, 5-32.

48. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* **2014***, 4*, 468-481.

49. Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000***, 40*, 185-194.

50. Dalgaard, P. *Introductory statistics with R;* Springer Science & Business Media: 2008.

51. Walker, T.; Grulke, C. M.; Tropsha, A.; Pozefsky, D. Chembench: A cheminformatics workbench. *Bioinformatics* **2010**, *26*, 3000-3001.

52. Kim, M. T.; Sedykh, A.; Chakravarti, S. K.; Saiakhov, R. D.; Zhu, H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. *Pharm. Res.* **2014**, *31*, 1002-1014.

53. Solimeo, R.; Kim, M.; Zhu, H.; Zhang, J.; Sedykh, A. Predicting chemical ocular toxicity using a combinatorial QSAR approach. *Chem. Res. Toxicol.* **2012**, *25*, 2763-2769.

54. Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913-1921.

55. Zhang, J.; Zhu, H.; Hsieh, J. H. Profiling animal toxicants by automatically mining public bioassay data: A big data approach for computational toxicology. *PLoS ONE* **2014**, *9*, 1-11.

56. Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **2008**, 506-513.

57. Wang, W.; Kim, M.; Sedykh, A.; Zhu, H. Developing enhanced blood-brain barrier permeability models: Integrating external bio-assay data in QSAR modeling. *Pharm. Res.* **2015**, *32*, 3055-3065.

58. Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ. Health Perspect.* **2011**, *119*, 364-370.

59. Métivier, R.; Penot, G.; Hübner, M.,R.; Reid, G.; Brand, H.; Kos, M.; Gannon, F. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **2003**, *115*, 751-763.

60. Purcell, D. J.; Jeong, K. W.; Bittencourt, D.; Gerke, D. S.; Stallcup, M. R. A Distinct Mechanism for Coactivator versus Corepressor Function by Histone Methyltransferase G9a in Transcriptional Regulation. *J. Biol. Chem.* **2011**, *286*, 41963-41971.

61. Tsuchiya, Y.; Nakajima, M.; Yokoi, T. Cytochrome P450-mediated metabolism of estrogens and its regulation in human. *Cancer Lett.* **2005**, *227*, 115-124.

62. Patlewicz, G.; Ball, N.; Becker, R. A.; Booth, E. D.; Cronin, M. T. D.; Kroese, D.; Steup, D.; Van Ravenzwaay, B.; Hartung, T. Read-across approaches - Misconceptions, promises and challenges ahead. *Archivos de Medicina Veterinaria* **2014**, *46*, 387-396.