

Classification and Virtual Screening of Androgen Receptor Antagonists

Jiazhong Li^{†,‡} and Paola Gramatica^{*,†}

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, via Dunant 3, 21100 Varese, Italy and Department of Chemistry, Lanzhou University, Tianshui South Road 222, 730000 Lanzhou, China

Received February 24, 2010

Computational tools, such as quantitative structure–activity relationship (QSAR), are highly useful as screening support for prioritization of substances of very high concern (SVHC). From the practical point of view, QSAR models should be effective to pick out more active rather than inactive compounds, expressed as sensitivity in classification works. This research investigates the classification of a big data set of endocrine-disrupting chemicals (EDCs)—androgen receptor (AR) antagonists, mainly aiming to improve the external sensitivity and to screen for potential AR binders. The *k*NN, lazy IB1, and ADTree methods and the consensus approach were used to build different models, which improve the sensitivity on external chemicals from 57.1% (literature) to 76.4%. Additionally, the models' predictive abilities were further validated on a blind collected data set (sensitivity: 85.7%). Then the proposed classifiers were used: (i) to distinguish a set of AR binders into antagonists and agonists; (ii) to screen a combined estrogen receptor binder database to find out possible chemicals that can bind to both AR and ER; and (iii) to virtually screen our in-house environmental chemical database. The *in silico* screening results suggest: (i) that some compounds can affect the normal endocrine system through a complex mechanism binding both to ER and AR; (ii) new EDCs, which are nonER binders, but can *in silico* bind to AR, are recognized; and (iii) about 20% of compounds in a big data set of environmental chemicals are predicted as new AR antagonists. The priority should be given to them to experimentally test the binding activities with AR.

1. INTRODUCTION

Many environmental compounds interfering with the body's endocrine system, which produce adverse developmental, reproductive, neurological, and immune effects in both human and wildlife,¹ are named as endocrine-disrupting chemicals (EDCs). A wide range of substances, both natural and man-made, are thought to cause endocrine disruption, including pharmaceuticals, dioxins and dioxin-like compounds, polychlorinated biphenyls, DDT and other pesticides, and plasticizers, such as bisphenol A, etc. The underlying mechanism of the EDCs can be summarized as two main categories. One is the inhibition of the biosynthesis or metabolism of endogenous ligands to indirectly modulate endocrine function (nonreceptor-mediated disruptors).² Another one is the direct interaction of a chemical with target steroid hormone receptors to interfere with the ligand-dependent transcriptional function (receptor-mediated disruptors).³ Receptor ligands can be classified into agonist and antagonist. An agonist binds to a receptor of a cell and triggers a response by the cell—mimicking the action of a naturally occurring substance. An antagonist acts opposite to an agonist and blocks the receptor activation. In competitive antagonism, binding of the antagonist to the receptor prevents binding of the agonist to the same receptor. In noncompetitive antagonism, antagonist and agonist can be bound simultaneously, but antagonist binding reduces the action of the agonist.⁴

The endocrine hormone receptors are a large family of ligand-dependent transcriptional factors known as the steroid receptor superfamily,^{5,6} such as estrogen receptor (ER) and androgen receptor (AR). The study about chemicals interacting with ER has been the focus of research for more than 20 years.^{7–12} Comparatively, the study about AR disruptors was just started more recently. Though some structurally diverse chemicals that can bind to and affect transactivation of AR have been recognized, such as steroids, synthetic hormones, polycyclic aromatic hydrocarbons (PAHs), etc.,^{13,14} there are more and more environmental chemicals which need to be tested and identified. Limited testing resources have led to a call for alternative and faster methods of screening AR disruptors with sufficient accuracy. Quantitative structure–activity relationship (QSAR) technique shows its benefits on this challenge. Instead of arduous and expensive laboratory work, biological activity can be predicted solely based on the molecular structure, decreasing the number of animal tests. The development and use of QSAR to assess the hazard of substances is expressly promoted and included in the REACH regulation¹⁵ not only to fill the data gap but also to design new chemicals for a progressive substitution of dangerous substances with suitable, safer substances, as required by the authorization step for the substances of very high concern (SVHC), such as EDCs. In the practical applications for prioritizing compounds to be tested experimentally, it could be sufficient to have a simple classification scheme that divides compounds into active and inactive.¹⁶

* Corresponding author e-mail: paola.gramatica@uninsubria.it.

[†] University of Insubria.

[‡] Lanzhou University.

Many QSAR researches have been done on chemicals with estrogen-like activities but very few on androgen-like activities mainly for two reasons: (i) AR disruptors cover a wider range of chemical structures, so it is difficult to develop a good general model; and (ii) there are very few experimental data on AR. So far, there are relatively few works on the screening of AR activity for large number of chemicals.^{2,7,13,14,17–19} Among these, an antagonist data set reported by Vinggaard et al.¹³ is the largest one, and this is also the first research to present AR transactivation data for chemicals with various skeleton structures and different functions. In their work, the antagonistic activities (human AR) of 292 compounds were first tested in a sensitive luciferase reporter assay,^{20,21} together with some chemicals collected from the literature, to generate the training data. Then the MultiCASE²² system was employed to develop a QSAR model. The model was used to externally predict 102 new chemicals, which were also tested by using the same assay, with a sensitivity of 57.1%, a specificity of 98%, and a concordance of 92%. The external concordance of this model is very high. But the high concordance value is due to the high specificity, while the sensitivity is very low. However, it is important to highlight that sensitivity is the most important parameter in a classification task. In fact, the low sensitivity value indicates the low ability of a model to recognize the active samples from the universe of chemicals, so it cannot provide sufficient accuracy particularly to distinguish molecules that should have the highest priority for animal test, being the more dangerous.

Though it is more difficult to predict active compounds than to predict inactive samples,²³ from the practical point of view, the QSAR model should recognize active samples as much as possible according to the precautionary principle. In this study, we try to develop QSAR classification models on the big AR antagonists data set with higher sensitivity than the published model.¹³ The same training and prediction sets were used.

It is reported that the poor external predictive power of the QSAR models for new chemicals could be due to the incorrect usage or lack of external validation during the modeling process.²⁴ After analyzing the distribution of the training and test samples, we found that their test set was so unbalanced between active and inactive compounds, thus we decided to combine all the data together and tried to split the data into more balanced and representative training and prediction sets. The *k*-nearest neighborhood (*k*NN),²⁵ local lazy IB1²⁶ and ADTree²⁶ methods, using DRAGON²⁷ descriptors, were employed to build different classification models based on the new training and prediction sets. The consensus approach was also used. Model predictive abilities were further validated on a new external data set collected from the literature,^{14,17–19} containing 89 highly heterogeneous compounds. Then the proposed classifiers were used: (i) to distinguish a set of AR binders reported by Fang et al.¹⁴ into antagonists and agonists, (ii) to screen a combined ER binder database^{28,29} to find out possible EDCs that can bind to AR or even to both AR and ER, and (iii) to virtually screen our wide in-house environmental chemical database.

2. MATERIALS AND METHODS

2.1. Data Sets. The studied data set was taken from the literature.¹³ The training set (training 1) included 292 experimental data measured in a Danish lab and 231 chemicals collected from the papers. The experimental values expressed the ability (IC₂₅, μ M) of a chemical to inhibit the luminescence response induced by the synthetic androgen, R1881. If the chemical reached an IC₂₅ at a test concentration $\leq 10 \mu$ M, then the chemical was defined as “active”. If an IC₂₅ was not reached or if the chemical gave rise to cytotoxicity at concentrations $\geq 3 \mu$ M, then the chemical was defined as “inactive”. Another 102 experimental data tested in the same lab were used as the external prediction set (test 1), including 14 active and 88 inactive compounds. All these 625 compounds together with corresponding Chemical Abstract Services (CAS) numbers (structures available from PubChem)³⁰ and experimental assignments (A, active or I, inactive) are listed in the Supporting Information, Table SII.

As discussed in Section 3.2.1, test 1 was so unbalanced between active and inactive samples, and the representativity of its active samples was not good enough to properly evaluate the model's predictive ability. Thus we resampled all the 625 chemicals and split them into more balanced and representative training set (training 2) and prediction set (test 2) by using Kohonen self-organizing mapping (SOM)^{31,32} (17*17 neurons, 500 epochs). SOM takes advantage of the clustering capabilities, allowing the selection of a meaningful training set and a representative prediction set. After training, similar chemicals fall within the same neuron, i.e., they carry similar information. To select the training set of chemicals, it is assumed that the compound closest to each neuron centroid is the most representative of all the chemicals within the same neuron. Thus, the selection of the training set chemicals was performed by the minimal distance from the centroid of each cell in the top map. The remaining objects, close to the training set chemicals, were used for the prediction set.







To further check the model predictive ability, 89 new compounds collected from different papers^{14,17–19} were used as a new external prediction set (test 3), listed in Table 1.

In addition, Fang et al.¹⁴ have reported 146 AR binders, but no distinction between agonist and antagonist was made in their paper. Out of those, 95 of them (listed in Table 2) were not included in the work of Vinggaard et al.¹³ Our models were used to distinguish these AR binders.

2.2. Databases for Virtual Screening. We screened two databases in this work. The first one is the combination of METI²⁸ and EDKB,²⁹ where the chemicals are experimentally tested as ER binder or nonbinders. It is reported that some typical ER ligands (e.g., DES, steroidal estrogens, and bisphenol A) are also active in AR binding,¹⁴ and generally they possess estrogenic and antiandrogenic activities.^{2,7,18,19} This kind of chemicals must affect the health of humans and wildlife through complex mechanisms. More research is needed to understand their action with the endocrine systems.

Both of these databases contain determined values of human ER. METI²⁸ is one of the largest data collections publicly available with more than 900 compounds. EDKB²⁹ is another big ER binder database. From EDKB, 87 chemicals have experimental values determined with human ER. Some compounds appear in both databases. If the

Table 1. Compounds Contained in Test 3 and Corresponding Experimental and Predicted Assignments

No	CAS	Class	kNNb	Lazy	ADTree	Consensus	Ref.	No	CAS	Class	kNNb	Lazy	ADTree	Consensus	Ref.
1	101-61-1	I	I	I	A	I	14	51	22350-76-1	I	A	A	A	A	18
2	104-51-8	I	I	I	I	I	14	52	2595-54-2	I	I	I	I	I	18
3	108-95-2	I	I	I	I	I	14	53	319-84-6	I	I	I	I	I	18
4	1117-86-8	I	I	I	I	I	14	54	319-85-7	I	A	I	I	I	18
5	119-36-8	I	I	I	I	I	14	55	33089-61-1	I	I	I	A	I	18
6	121-33-5	I	I	A	I	I	14	56	3811-49-2	I	I	A	I	I	18
7	124-20-9	I	I	I	I	I	14	57	3861-47-0	I	I	A	I	I	18
8	135-98-8	I	I	I	I	I	14	58	391-86-8	I	I	I	I	I	18
9	140-10-3	I	I	I	I	I	14	59	41814-78-2	I	I	I	I	I	18
10	143-74-8	I	A	I	A	A	14	60	5103-71-9	A	A	A	A	A	18
11	145-13-1	I	A	I	A	A	14	61	5103-74-2	A	A	A	A	A	18
12	15372-34-6	I	A	I	A	A	14	62	533-74-4	I	I	I	I	I	18
13	1833-27-8	I	I	I	A	I	14	63	55814-41-0	I	A	I	A	A	18
14	2385-85-5	I	I	I	I	I	14	64	56-38-2	A	I	I	A	I	18
15	3391-86-4	I	I	I	I	I	14	65	57369-32-1	I	I	I	I	I	18
16	50-24-8	I	A	A	A	A	14	66	57837-19-1	I	I	I	I	I	18
17	505-48-6	I	I	I	I	I	14	67	58858-18-7	I	A	I	A	A	18
18	52-39-1	I	A	A	A	A	14	68	60-51-5	I	I	I	I	I	18
19	539-08-2	I	A	A	I	A	14	69	61432-55-1	I	I	I	I	I	18
20	57-10-3	I	I	I	I	I	14	70	66332-96-5	I	I	A	A	A	18
21	58-08-2	I	I	I	I	I	14	71	68085-85-8	I	I	I	A	I	18
22	59-30-3	I	I	I	A	I	14	72	68505-69-1	I	I	A	A	A	18
23	629-41-4	I	I	I	I	I	14	73	6923-22-4	I	I	I	I	I	18
24	65-45-2	I	I	I	I	I	14	74	69409-94-5	I	I	I	I	I	18
25	6665-86-7	I	I	I	I	I	14	75	72-20-8	I	A	A	I	A	18
26	73-31-4	I	I	I	I	I	14	76	7292-16-2	I	I	I	A	I	18
27	81-90-3	I	A	A	A	A	14	77	85785-20-2	I	I	I	I	I	18
28	84-60-6	I	I	A	I	I	14	78	87130-20-9	I	I	I	I	I	18
29	886-65-7	I	A	A	A	A	14	79	92-52-4	I	I	I	I	I	18
30	94871-36-0	I	I	I	A	I	14	80	959-98-8	A	A	A	I	A	18
31	95-57-8	I	I	I	I	I	14	81	13171-00-1	I	A	A	I	A	19
32	140-66-9	A	I	A	A	A	17	82	496-11-7	I	I	I	I	I	19
33	16561-29-8	I ^a	A	A	I	A	17	83	13311-84-7	A	A	A	A	A	19
34	521-18-6	I ^a	I	A	I	I	17								
35	58-22-0	I	I	I	I	I	17								
36	68047-06-3	I ^a	A	A	A	A	17	84		A	A	A	A	A	19
37	76-43-7	I	I	I	A	I	17								
38	965-93-5	I	A	A	I	A	17	85		I	I	I	I	I	19
39	2164-08-1	I	I	I	I	I	18								
40	2439-01-2	I	I	I	I	I	18	86		I	I	I	I	I	19
41	2957-03-7	I	I	I	I	I	18								
42	1071-83-6	I	I	I	I	I	18	87		I	A	A	A	A	19
43	115-90-2	I	I	I	I	I	18								
44	121-75-5	I	I	I	I	I	18	88		I	A	A	A	A	19
45	13067-93-1	I	I	I	A	I	18								
46	13071-79-9	I	I	A	I	I	18	89		I	I	A	I	I	19
47	133220-30-1	I	I	I	A	I	18								
48	137-26-8	I	A	A	I	A	18								
49	1582-09-8	I	I	I	I	I	18								
50	22224-92-6	I	I	I	A	I	18								

^a Anticipated negative in ref 17.

duplicates belong to the same class (active or inactive), then one of them is deleted from the combined data set. If their assignments are inconsistent, then we erased both of them in order to avoid any noise from the experimental data. After further deleting the ionic compounds and those appearing in AR data set, we get the final data set containing 708 chemicals, listed in the Supporting Information, Table SI2.

Another data set is our in-house environmental database. We have studied environmental chemicals for many years in our lab. So far we have collected an environmental database containing 2767 chemicals of environmental concerns, including heterogeneous chemicals, such as polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), aromatic amines, plastic additives, pesticides, etc. Most of them are widespread in our environment and are harmful to the health of humans and wildlife.

2.3. Molecular Descriptors. The two-dimensional (2D) molecular structures were downloaded from PubChem.³⁰ All these structures were carefully verified and optimized by using the semiempirical AM1 method to the minimum energy

conformations in the HyperChem³³ program. The obtained conformations were submitted to the DRAGON²⁷ package to calculate 2914 theoretical descriptors, including: (a) 0D constitutional descriptors, (b) 1D count of functional groups and atom-centered fragments; (c) 2D topological descriptors, walk and path counts, connectivity and information indexes, various autocorrelations from the molecular graph, edge adjacency indices, descriptors of Burden eigenvalues,^{34,35} topological charge and eigenvalues-based indices, and 2D binary and frequency fingerprints; (d) 3D Randic molecular profiles, geometrical descriptors, weighted holistic invariant molecular descriptors (WHIMs),³⁶ and geometry, topology, and atom-weights assembly (GETAWAY) descriptors,³⁷ (e) charge descriptors, and (f) molecular properties. The list and meaning of the molecular descriptors is provided by the DRAGON package, and the calculation procedure is explained in detail, with related references, in a Handbook of Molecular Descriptors.³⁸

The constant or near-constant descriptors were deleted in a prerelation step. If the pairwise correlation of two

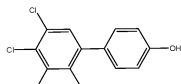
descriptors was very high (defined by the user, a K correlation coefficient greater than 0.95 here), then the one showing the highest pair correlation with all the other descriptors was automatically excluded. Finally 855 descriptors underwent a subsequent variable selection process.

2.4. Descriptor Selection. The support vector machines³⁹ recursive feature elimination (SVM-RFE) method, proposed by Guyon et al.,⁴⁰ was used in this research as the feature selection method for selecting molecular descriptors associated to AR antagonist activity. RFE is an iterative procedure for backward feature elimination, which can be executed in WEKA,²⁶ where the descriptors are normalized by default. The first step of SVM-RFE is to train a SVM classifier with all the features. Suppose there is a training set, $x = (x_1, x_2, \dots, x_i, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_i, \dots, y_n)$. SVMs solve the classification problem by minimizing the following equation:

$$J = (1/2) \sum_{ij} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j + \lambda \delta_{ij}) - \sum_i \alpha_i \quad \text{subject to} \quad 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_i \alpha_i y_i = 0$$

where α is the Langrange coefficient, δ_{ij} is the Kronecker symbol (if $i = j$ and $\delta_{ij} = 1$, otherwise 0), λ and C are two parameters needed to be optimized by SVM. The output of this solution is α_i . The resulting decision function is: $f(x) = w \cdot x + b$, where w is the weight vector calculated as $w = \sum_i \alpha_i y_i x_i$. The ranking criterion is the square of the weight, calculated as $c_l = (w_l)^2$ for the l th feature. Then the feature with the smallest c_l is removed. For computational reasons, it may be more efficient to remove several features in each cycle. At the end, all the features are ranked. The features on the top are the more informative ones.

Table 2. AR Binders Reported by Fang et al.¹⁴ and Corresponding Predictions (Agonist, G, or Antagonist, T)

No	CAS	Agonist (G) or antagonist (T)				No	CAS	Agonist (G) or antagonist (T)			
		kNNb	Lazy	ADTree	Consensus			kNNb	Lazy	ADTree	Consensus
1	101-92-8	T	T	G	T	50	50-23-7	T	T	T	T
2	103-16-2	G	G	T	G	51	5108-94-1	T	T	T	T
3	104-43-8	G	G	G	G	52	521-17-5	G	G	T	G
4	10540-29-1	G	T	T	T	53	525-82-6	T	T	G	T
5	1057-07-4	G	G	G	G	54	53-16-7	T	G	T	T
6	105-99-7	G	G	G	G	55	53-19-0	T	T	T	T
7	108-43-0	G	G	G	G	56	53-41-8	G	T	G	G
8	1137-42-4	T	T	T	T	57	53-43-0	G	T	T	T
9	115-29-7	T	T	G	T	58	53-63-4	T	G	T	T
10	1156-92-9	G	T	T	T	59	552-80-7	T	T	T	T
11	115-86-6	T	T	T	T	60	56-33-7	T	T	T	T
12	117-84-0	G	G	G	G	61	564-35-2	T	G	G	G
13	119-61-9	T	G	T	T	62	571-20-0	G	T	G	G
14	1224-92-6	G	T	G	G	63	571-22-2	G	G	G	G
15	1225-43-0	G	T	G	G	64	57-91-0	T	G	T	T
16	13026-26-1	T	T	T	T	65	58-72-0	G	G	T	G
17	13037-86-0	G	G	G	G	66	5975-78-0	G	T	T	T
18	13049-13-3	T	T	T	T	67	5976-61-4	T	T	T	T
19	131-56-6	T	T	T	T	68	599-64-4	T	T	T	T
20	141-04-8	G	G	G	G	69	603-45-2	T	T	T	T
21	1482-70-8	T	G	G	G	70	611-99-4	T	T	T	T
22	14868-03-2	T	T	T	T	71	62133-07-7	G	G	G	G
23	1570-64-5	T	T	G	T	72	6515-37-3	G	G	T	G
24	15872-42-1	G	G	T	G	73	659-22-3	T	T	T	T
25	1806-26-4	G	G	G	G	74	6665-83-4	T	T	T	T
26	1845-11-0	T	T	T	T	75	68-22-4	T	G	G	G
27	1852-53-5	G	T	G	G	76	68-23-5	T	G	G	G
28	20426-12-4	G	G	T	G	77	71030-11-0	T	T	T	T
29	2132-70-9	T	T	T	T	78	74738-17-3	T	T	T	T
30	2142-01-0	G	G	G	G	79	75938-34-0	T	T	T	T
31	24124-24-1	G	G	T	G	80	77-40-7	T	T	T	T
32	2437-79-8	T	T	T	T	81	791-31-1	T	T	G	T
33	25154-52-3	G	G	G	G	82	79199-51-2	T	T	T	T
34	2529-64-8	T	T	G	T	83	797-63-7	T	G	G	G
35	2657-25-2	G	G	T	G	84	80-09-1	T	T	G	T
36	2971-36-0	T	T	T	T	85	80-46-6	T	G	T	T
37	330-55-2	T	T	G	T	86	84-69-5	T	G	G	G
38	33330-65-3	T	G	T	T	87	89-72-5	G	G	G	G
39	34184-77-5	T	T	G	T	88	911-45-5	T	G	T	T
40	3434-79-5	T	T	T	T	89	92-69-3	T	T	T	T
41	362-05-0	T	T	T	T	90	94-25-7	G	G	G	G
42	36455-72-8	G	G	T	G	91	94-41-7	G	T	G	G
43	3704-09-4	G	T	G	G	92	97-54-1	G	G	T	G
44	3839-46-1	T	T	T	T	93	98-54-4	T	T	G	T
45	4180-23-8	G	G	T	G	94	99-71-8	T	G	T	T
46	42422-68-4	G	G	T	G	95		T	T	T	T
47	4250-77-5	G	G	G	G						
48	438-22-2	G	G	T	G						
49	487-26-3	T	T	G	T						

2.5. Modeling Methods. In this research, three classification methods were used, i.e., k -nearest neighbors (k NN), local lazy method (lazy IB1), and alternating decision tree (ADTree).

2.5.1. k -Nearest Neighbors (k NN). The k NN method⁴¹ is a simple classification method based on local information around each object. k NN is a nonparametric method where the classification of an object depends on the class assignments of its k -nearest neighbors, without making any assumptions about the distribution and the shape of the classes or about the form of class boundaries. The nearness is measured by an appropriate distance metric (e.g., Euclidean distance). The standard k NN method^{41,42} is implemented simply as follows: (i) calculate distances between each unknown object (u) and all the objects in the training set; (ii) select a range for k ; (iii) for each k value, the class to which a majority of the k -nearest training objects belong is assigned to each query u ; (iv) the k value giving the lowest leave-one-out (LOO) cross-validation error rate is the optimal and is used for new object prediction.

2.5.2. Local Lazy Method (Lazy IB1). Lazy learners is a memory-instance-based learning technique, which stores the training objects and does no real work until a prediction is required for an unknown object (u). The term lazy arises because the predictions for the test set compounds are made without producing a model a priori on the whole training set. Considering the close neighborhood of a query point according to a distance (Euclidean distance here) measure, the activity of the query is predicted from the activities of the most chemically similar neighbor compounds in the training set. Lazy IB1, which can be executed in WEKA,²⁶ is a basic learner. Once the nearest training sample has been located, IB1 predicts the same class as the training sample for u . If several samples qualify as the closest, then the first one found is used.

2.5.3. Alternating Decision Tree (ADTree). Alternating decision trees (ADTree)⁴³ is a kind of option tree. ADTree has been used as a tool to mine the NCI human tumor cell line database⁴⁴ and to analyze the mass spectrometry data,⁴⁵ etc. Option trees differ from decision trees in that they contain two types of nodes: a decision node and a prediction node, while decision trees just contain a decision node. When a query reaches a decision node, the sign of this node will be assigned to the query, like in the decision tree. However, when the query reaches a prediction node, it will continue to all the paths of this node. So in an alternating decision tree, the studied compound could follow different branches (multipath). The sign of the sum of all the prediction nodes that is included in a multipath is the class which the tree associates to the query. One possibility to grow an option tree is incrementally adding nodes to a decision tree. This is commonly done by using the boosting algorithm, and the resulted trees are usually called ADTree instead of option trees. The number of boosting iterations is an important parameter that can be tuned to suit the data set and the desired complexity–accuracy trade-off, which was set as 20 in this work. The default search method of exhaustive search (expands all paths) in WEKA was used in this research.

2.6. Model Evaluation. Sensitivity and specificity are common statistical measures of the performance for a binary classification research. The sensitivity measures the proportion of actual positives which are correctly identified; and

the specificity measures the proportion of negatives which are correctly identified, defined as follows:

$$\text{Sensitivity} = 100\text{TP}/(\text{TP} + \text{FN})$$

where TP means true positive, and FN means false negative.

$$\text{Specificity} = 100\text{TN}/(\text{TN} + \text{FP})$$

where TN is true negative, and FP is false positive.

A theoretical, optimal prediction can achieve both 100% sensitivity and specificity. For any test, there is usually a trade-off between each measure. In the research where one is testing for hazard, such as androgen antagonist in this work, one may be willing to risk discarding functioning components (low specificity), in order to increase the chance of identifying nearly all dangerous components (high sensitivity). So in this research, sensitivity is the most important parameter, which indicates the ability to recognize the active compounds (AR antagonists) from a universe of chemicals. These two parameters together with the concordance (total accuracy) were employed to estimate the performance of the built classification models:

$$\text{Concordance} = 100(\text{TN} + \text{TP})/(\text{TN} + \text{TP} + \text{FN} + \text{FP})$$

In addition, a receiver operating characteristic (ROC) curve was employed to graphically present the model behavior in a visual way. A ROC curve, which has been proved to be a valuable way to evaluate the quality of a two-class classifier, shows the separation ability of a binary classifier by iteratively setting the possible classifier threshold. As a result, a plot of the trade-off between the sensitivity (y -axis) and 1-specificity (x -axis) is shown. If the plot has a surface area of 1, a perfect classifier is found, and if the area equals 0.5, the classifier has no discriminative power at all.

2.7. Applicability Domain (AD). To verify the practical applicability of our models to chemicals not used in model development, the model's applicability domain, which is a theoretical region defined by the used descriptors in modeling, was quantitatively assessed by the leverage approach.^{46–48} The leverage (h) is calculated by $h_i = x_i(X^T X)^{-1} X_i^T$ ($i = 1, \dots, m$), where x_i is the descriptor row vector of the query compound i , m is the number of query compounds, and X is the $n \times k$ matrix of the training set (k is the number of model descriptors, and n is the number of training set samples). The limit of normal values for X outliers (h^*) is set as $3(k + 1)/n$, and a leverage greater than h^* means that the prediction is the result of substantial extrapolation of the model and could not be reliable.

3. RESULTS AND DISCUSSION

3.1. Classification Model Based on Training 1 (k NNa). This study intends to develop QSAR models with high external sensitivity. At first, the same training data as used in literature¹³ (training 1) was used to build a new QSAR model to check the utility of our methods on this data set. After descriptors calculation and prereluction in DRAGON, the RFE method was used to rank all the descriptors. If one descriptor has nonzero values for very few compounds (typically counts or finger prints), it was deleted among the rank list. Then k NN models are built with the top descriptors, and the number of descriptors is

Table 3. The Descriptors Selected Basing on Training 1 and Their Corresponding Meanings

descriptor	meaning	descriptor type
VEA1	eigenvector coefficient sum from adjacency matrix	2D eigenvalue-based indices
EEig14d	eigenvalue 14 from edge adj. matrix weighted by dipole moments	2D edge adjacency indices
B02[N–N]	Presence/absence of N–N at topological distance 2	2D binary fingerprints
G1e	The first component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities	3D WHIM descriptors
F05[O–Cl]	Frequency of O–Cl at topological distance 5	2D frequency fingerprints
BEHp2	Highest eigenvalue <i>n</i> . 2 of Burden matrix/weighted by atomic polarizabilities	2D Burden eigenvalue descriptors
F08[C–N]	Frequency of C–N at topological distance 8	2D frequency fingerprints
FDI	Folding degree index	3D geometrical descriptors
BEHe3	Highest eigenvalue <i>n</i> . 3 of Burden matrix/weighted by atomic Sanderson electronegativities	2D Burden eigenvalue descriptors
SPAM	average span <i>R</i>	3D geometrical descriptors

Table 4. The Statistics of Different Classification Models

data set	parameters	model using training 1		model using training 2			
		MultiCASE*	kNNa	kNNb	lazy	ADtree	consensus
training sets	sensitivity (%)	67.9	74.4	69.7	67.2	64.7	72.6
	specificity (%)	75.5	74.0	79.3	76.6	73.6	82.7
	concordance (%)	72.1	74.2	75.4	72.8	70.0	78.6
	area under ROC curve	0.716	0.742	0.745	0.719	0.691	0.777
test 1	sensitivity (%)	57.1	71.4	—	—	—	—
	specificity (%)	97.6	87.8	—	—	—	—
	concordance (%)	91.7	85.4	—	—	—	—
	area under ROC curve	0.774	0.796	—	—	—	—
test 2	sensitivity (%)	—	—	72.7	74.5	74.5	76.4
	specificity (%)	—	—	81.1	73.0	77.0	82.4
	concordance (%)	—	—	77.5	73.6	76.0	79.8
	area under ROC curve	—	—	0.769	0.738	0.758	0.794
test 3	sensitivity (%)	—	57.1	71.4	85.7	85.7	85.7
	specificity (%)	—	63.4	75.6	72.0	68.3	74.4
	concordance (%)	—	62.9	75.3	73.0	69.7	75.3
	area under ROC curve	—	0.603	0.735	0.788	0.770	0.801

* Results from ref 13.

incrementally increased. When adding new descriptors, it does not increase the LOO accuracy for the training set, the best model is obtained. As a result, 10 variables (listed in Table 3) were selected to construct the *k*NN model (*k*NNa) in SCAN.⁴⁹ The correlations of the used descriptors, listed in the Supporting Information Table SI3, show that the descriptors do not exhibit significant intercorrelation among themselves, so they catch different structural information. The *k*NN model with a neighbor number of 5 (*k* = 5) gave a very similar statistical performance with sensitivity, specificity, and concordance around 74% for the training set, based on the LOO procedure (summarized in Table 4).

Then this model was externally validated by 102 experimental data (test 1). The *k*NN model could correctly predict 10 of 14 active compounds (10/14) with a sensitivity of 71.4%, which was much higher than 57.1% from the literature.¹³ Though the specificity and the concordance were lower than those in the literature, the advantage of our model is that we could recognize more active chemicals (androgen antagonists) from the external chemicals. This is the most important ability of classification models and the main aim of our work.

3.2. Classification Models Based on Training 2.

3.2.1. Data Analysis and Splitting. There is a general agreement that QSAR models should be validated by using chemicals that are not involved in the model-building process^{47,48} and that the compositions of the training and

external prediction sets are of crucial importance.^{50–52} The best splitting must guarantee that the two sets are scattered over the whole area occupied by representative points in the descriptor space (representativity).^{48,50–52} Analysis of test 1 reveals that it is too unbalanced with a ratio of 10:90 between active and inactive samples. A principal component analysis (PCA) of molecular descriptors was performed to explore the structural chemical space for training 1 and test 1 and the distribution of chemicals. PCA, the simplest of the true eigenvector-based multivariate analyses, involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. In the PCA plot, shown in Figure 1, the active and inactive samples in the prediction set are marked by “triangle” and “star”, respectively. It is obvious that the chemical diversity of test 1 is limited, especially for the 14 active samples. Though this prediction set was designed to reflect the “true” ratio in nature,¹³ the unbalanced peculiarity and poor representativity could not properly assess the model’s predictive ability.

Thus, we decided to combine together all the data used by the Danish group¹³ and to divide the complete data set into more balanced and representative training and prediction sets using the SOM method.^{31,32} Totally 44 most significant principal components of molecular descriptors (the first two or three calculated from each block of DRAGON descriptors) were used as variables to build a Kohonen top map, shown

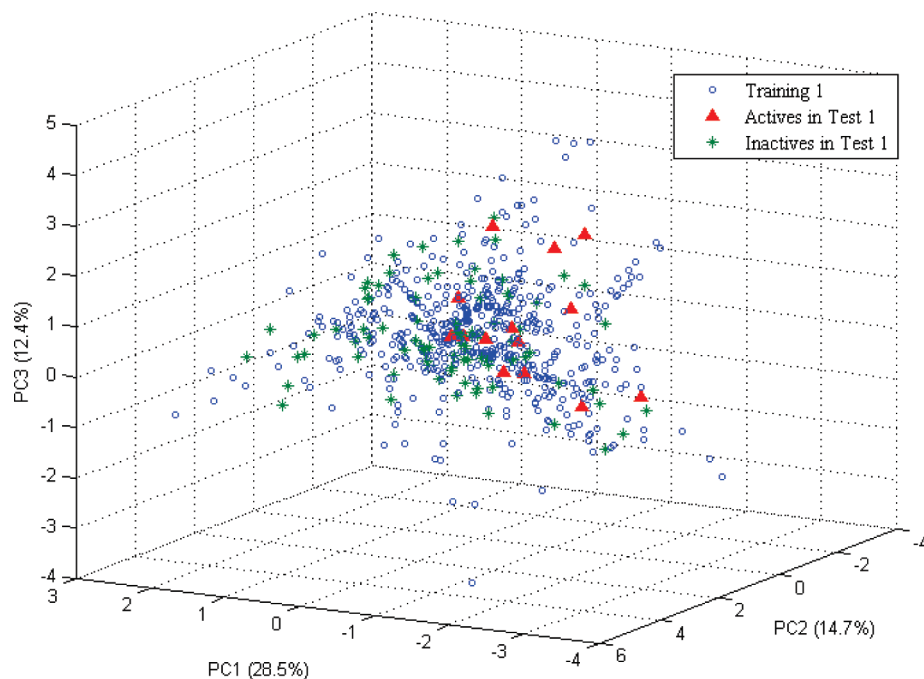


Figure 1. Distribution of training 1 and test 1 chemicals in the structural space, represented by the three most important principal components of molecular descriptors (cumulative explained variance = 55.6%).

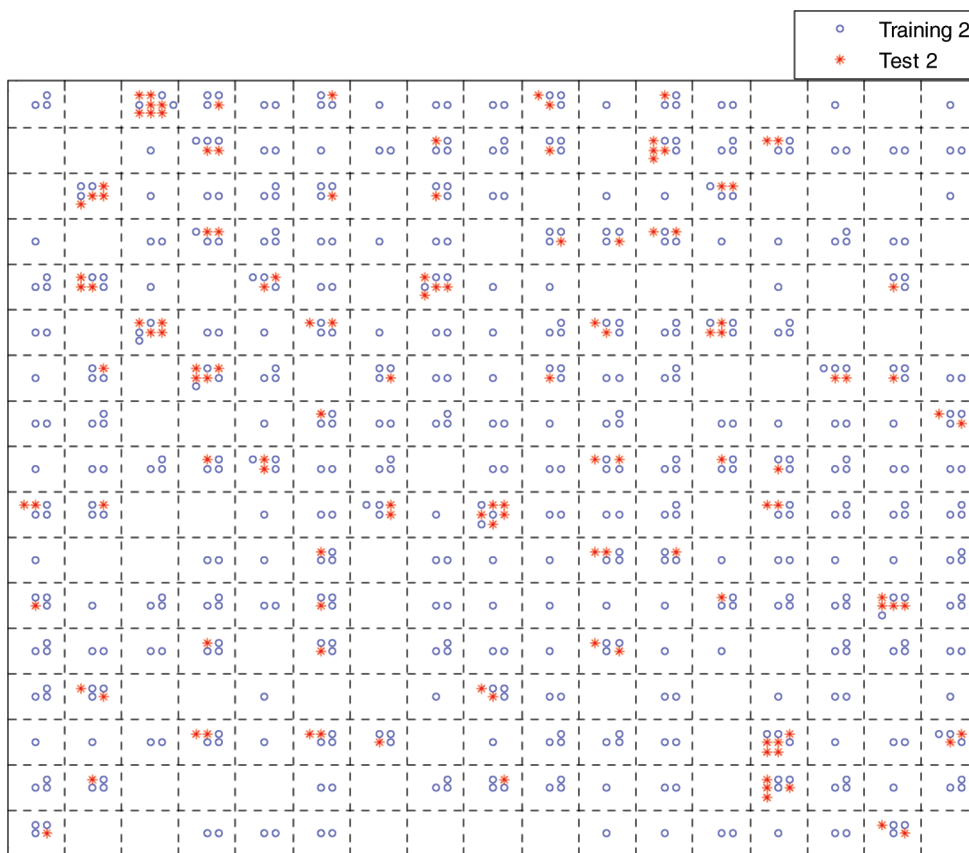


Figure 2. The top map for SOM splitting in training 2 and test 2.

in Figure 2. Chemicals with more similar structures were close to each other in the top map. The selection of the training set chemicals was performed by the minimal distance from the centroid of each cell. The remaining objects, close to the training set chemicals, were used as the prediction set.

After SOM clustering and splitting, 496 compounds were selected in the training set (training 2), and 129 compounds

fell into the prediction set (test 2), which were marked differently in Figure 2. The 3D distributions of these two data sets are shown in Figure 3, where the active compounds in test 2 are highlighted with a solid “triangle”. Comparing Figures 1 and 3, it is obvious that the samples in test 2 scatter over the whole descriptor space, especially the active compounds. The distribution of test 2 is much wider than

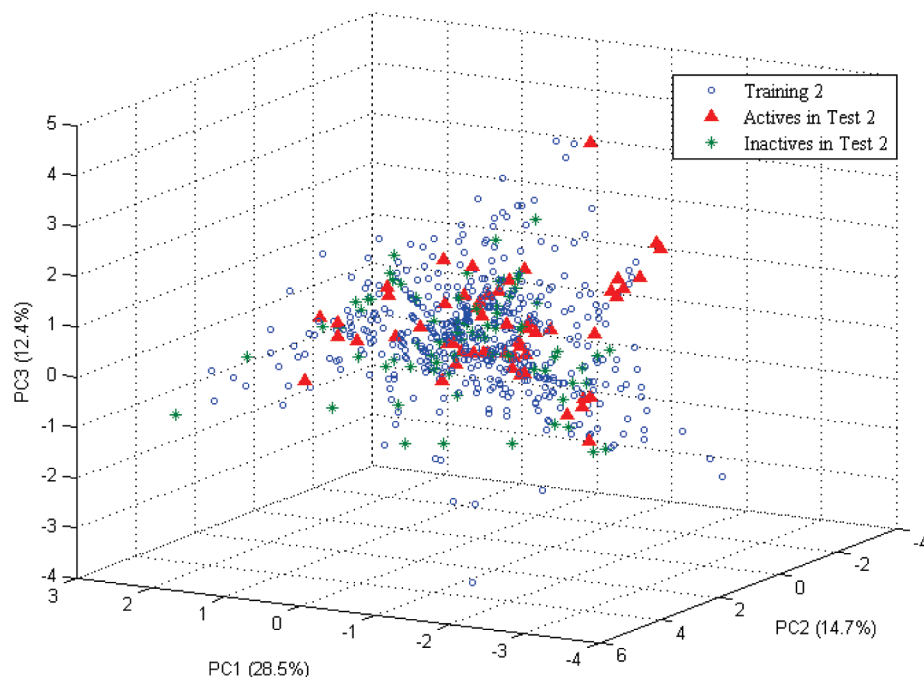


Figure 3. Distribution of training 2 and test 2 chemicals in the structural space, represented by the three most important principal components of molecular descriptors (cumulative explained variance = 55.6%).

Table 5. The Descriptors Selected Based on Training 2 and Their Corresponding Meanings

descriptor	meaning	descriptor type
BEHp2	highest eigenvalue n . 2 of Burden matrix/weighted by atomic polarizabilities	2D Burden eigenvalue
C-026	number of R-CX-R	1D atom-centered fragments
GATS7v	Geary autocorrelation — lag 7/weighted by atomic van der Waals volumes	2D autocorrelation indices
TPSA(NO)	topological polar surface area using N, O polar contributions	molecular properties
F08[C-O]	frequency of C-O at topological distance 8	2D frequency fingerprints
nRCONHR	Number of Y-NH-C(=O)-Al, where Y = Ar or Al (not H, not C=O) and Al = H or aliphatic group linked through C	1D functional group counts
H0e	H autocorrelation of lag 0/weighted by atomic Sanderson electronegativities	3D GETAWAY
FDI	folding degree index	3D geometrical
SPH	sphericity	3D geometrical
E2e	the second component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities	3D WHIM descriptors
R3u+	R maximal autocorrelation of lag 3/unweighted	3D GETAWAY

test 1, and test 2 has better representativity of the studied data set. Furthermore, the ratios of active/inactive are very similar between training 2 and test 2.

3.2.2. *kNN Model (kNNb) and Descriptor Explanation.*

3.2.2.1. *kNN Classification Model (kNNb).* RFE method was used to rank the descriptors based on training 2, and the top 11 descriptors in the list were used to build classification models. The corresponding physicochemical meanings are listed in Table 5. The selected descriptors are independent, as can be verified by their pair correlations in the Supporting Information, Table SI4. With the selected 11 descriptors, a *kNN* model (*kNNb*, $k = 4$) was established on the new training set giving an overall classification performances better than the literature and the *kNNA* model for the training set based on the LOO procedure (Table 4). Then this model was used to predict the new external prediction set (test 2), which includes 55 active compounds. The *kNNb* model can correctly predict 40 of them with a sensitivity of 72.7%, which is higher than *kNNA* and the published literature results.¹³ The large improvement of sensitivity comes

together with good specificity and concordance. The predicted assignments are given in the Supporting Information, Table SI1.

As described above, the ROC curve was employed to show the classification ability of the binary classifier. Figures 4 and 5 are the ROC plots for the training set and the external prediction set, respectively. In these two figures, the areas under the *kNNb* model lines for training 2 and test 2 are 0.745 and 0.769, respectively. The results indicated the high predictive ability of this model on external test chemicals.

3.2.2.2. *Descriptors Explanation.* Two kinds of structural information can be highlighted in two groups of selected descriptors: the folding degree index (FDI) and the sphericity (SPH) are molecular shape descriptors; BEHp2, E2e, H0e, TPSA(N-O), C-026, nRCONHR, and F08[C-O] catch different kinds of information about molecular polarity. A principal component analysis of chemicals in the space of these descriptors was executed, and the loading plot of the descriptors is shown in Figure 6.

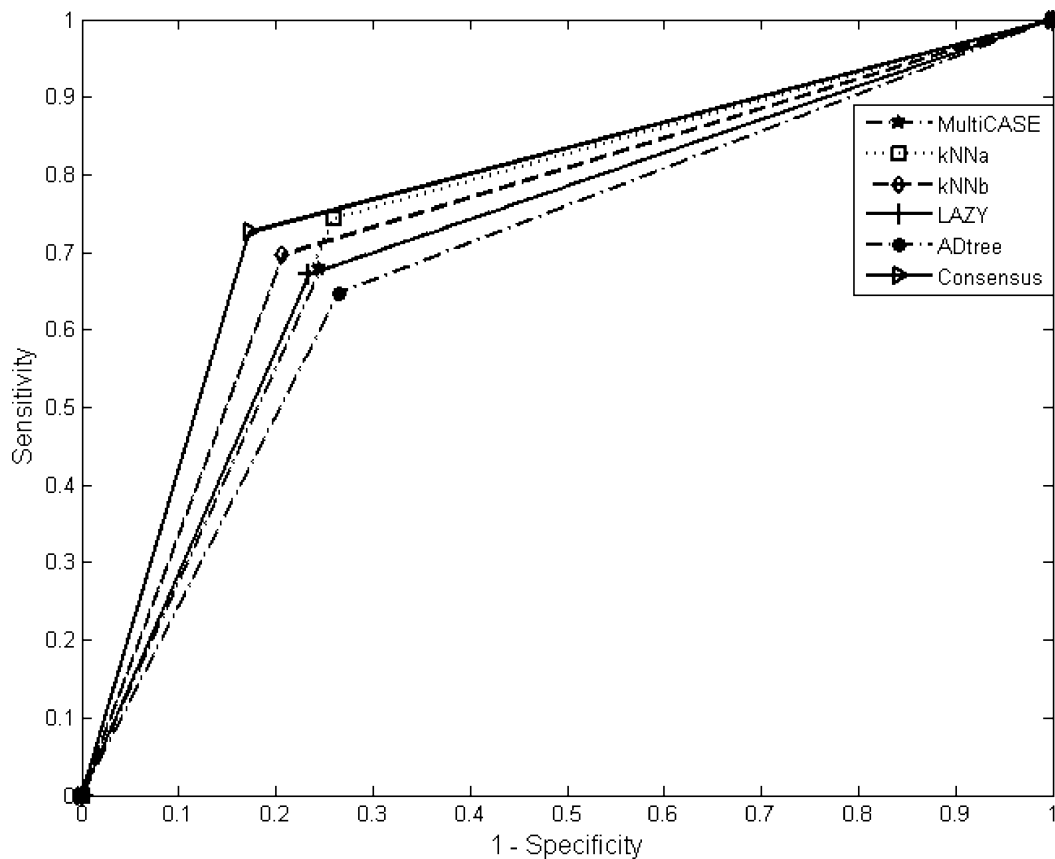


Figure 4. ROC plot on training set from different models. MultiCASE (from ref 13) and *k*NNa models are based on training 1. Other models are based on training 2.

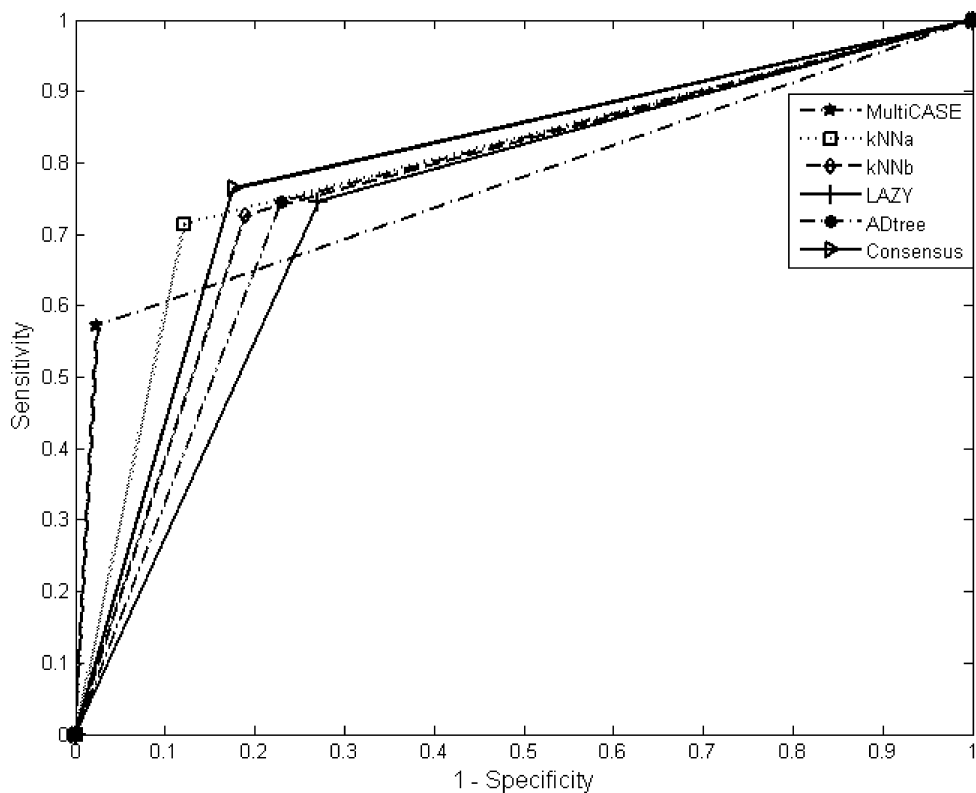


Figure 5. ROC plot on prediction set from different models. MultiCASE (from ref 13) and *k*NNa models are based on training 1. Other models are based on training 2.

In this figure, the FDI and SPH descriptors are both of high importance, positively contributing to principal component 1 (PC1). The folding degree index is the largest

eigenvalue of the distance/distance matrix normalized, dividing it by the number of atoms. This index tends to be one for linear molecules (of infinite length) and decreases in

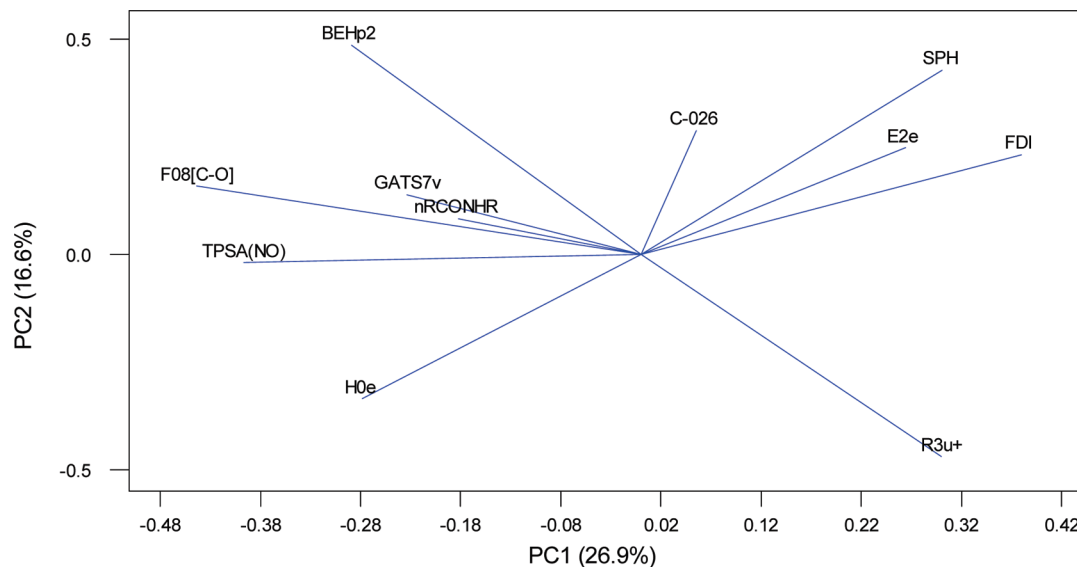


Figure 6. The principal components loading plot of the selected descriptors based on training 2. (EV%: 43.5%).

correspondence with the folding of the molecule. Thus, it can be treated as a measure of the molecular folding degree because it indicates the degree of distance from strict linearity. The sphericity (SPH), which is an anisometry descriptor calculated as a function of the eigenvalues of the covariance matrix, varies from 0 for flat molecules, such as benzene, to 1 for totally spherical molecules. We checked the compounds with high PC1 values in the same direction of FDI and SPH in Figure 6, almost all these compounds are simply substituted benzenes. In the opposite direction, bigger and more folded compounds are located.

BEHp2 is a Burden eigenvalue descriptor,^{34,35} indicating the highest eigenvalue n . 2 of Burden matrix/weighted by atomic polarizabilities. The Burden eigenvalue descriptors were originally proposed⁵³ to address searching for chemical similarity/diversity on large databases and are based on a significant extension of the Burden approach. Burden descriptors have been demonstrated to reflect relevant aspects (here polarizability) of molecular structure. E2e is the second component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities. WHIM descriptors³⁶ are built to capture relevant molecular 3D information regarding molecular size, shape, symmetry, and atom distribution (here taking into account their electronegativity). H0e is a GETAWAY descriptor³⁷ representing the H autocorrelation of lag 0, also weighted by atomic Sanderson electronegativities. These three descriptors are holistic, representing the electronegative properties of the whole molecule. Other four descriptors, TPSA(N-O), C-026, nRCONHR, and F08[C-O], are all fragment variables, representing the influence of a particular functional group, especially based on the presence of electronegative atoms. TPSA(NO) is the topological polar surface area using N, O polar contributions. F08[C-O] is the 2D fingerprint descriptor counting the distance between C and O atoms. It is easy to understand that the compounds in the direction of high values of TPSA(NO) and F08[C-O] in Figure 6 have more heteroatoms, such as N and O. C-026 and nRCONHR are the number of group R-CX-R (with X representing any electronegative atom O, N, S, halogens) and RCONHR, respectively.

The left two descriptors are GATS7v and R3u+. GATS7v is a 2D autocorrelation descriptor, which explains how the values of certain functions (here van der Waals volumes), at intervals equal to the lag d , are correlated. These descriptors can be obtained by summing up the products of certain properties of the two atoms located at a given topological distance or spatial lag. In general, they describe how the considered property is distributed along the topological structure. GATS7v indicates Geary autocorrelation — lag 7/weighted by atomic van der Waals volumes. R3u+ belongs to GETAWAY descriptors, specifically to R-GETAWAY. These descriptors come from the influence/distance matrix (R) where the elements of the molecular influence matrix are combined with those of the geometric matrix. R3u+ represents an R index of maximal contribution to the autocorrelation in lag 3 (topological distance).

It is important to keep in mind that in any multivariate study, such as in QSAR models which are based on several molecular descriptors, the studied activity (here the binding affinity for AR) is modeled not only by each single descriptor at time but by the combination of all of them. The discrimination ability of the classification models is based on the cumulative structural information caught simultaneously by different kinds of descriptors, thus it is highly difficult and too ambitious to link each selected descriptor to the observed activity. However, from the above discussion, we can verify that, in a holistic view, the molecular shape and size, and mainly the electronic distribution, due to the presence of electronegative atoms, and the derived molecular polarity are crucial factors for the AR binding ability.

3.2.3. Lazy IB1 Prediction. The same 11 descriptors were used in WEKA program to make the lazy IB1 prediction, and the LOO performances for the training 2 set are listed in Table 4. As shown in Figure 4, the ROC plot of the lazy prediction is under the k NNb curve for the training set with the area of 0.719, which is a little worse. As for the external prediction on test 2, lazy prediction gives a highest sensitivity of 74.5%. The lazy IB1 prediction curve is below k NNb curve in Figure 5 due to the lower specificity and concordance.

3.2.4. ADTree Classification Model. Then nonlinear ADTree method was used to develop a classification model based

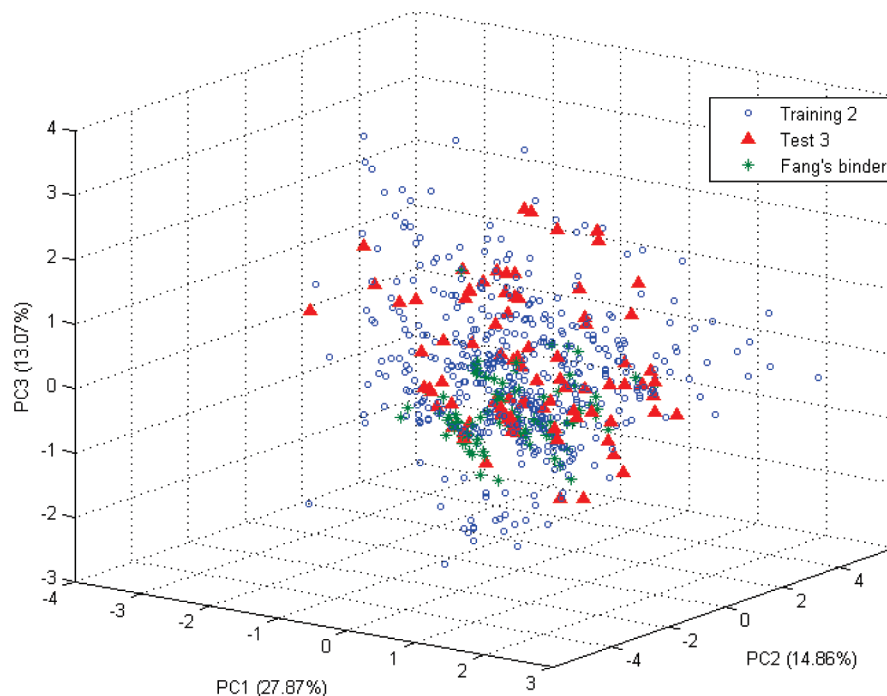


Figure 7. Distribution of test 3 and Fang's AR binder in training 2 structural space represented by the three most important principal components of molecular descriptors (cumulative explained variance = 55.8%).

on the same subset of descriptors as inputs. The predicted results were listed in the Supporting Information, Table SI1. In the ROC plot for the training set (Figure 4), the ADTree model curve is at the bottom with the smallest area of 0.691. Actually, the ADTree model has the lowest classification results (Table 4). For the external prediction set, this model performs better than the lazy prediction with a ROC plot area of 0.758 in Figure 5. These two models have the same sensitivities of 74.5%, but the specificity of 77% from ADTree model is much higher. The concordance for test 2 is 76%, similar to the *k*NNb model.

3.2.5. Consensus Analysis. The statistical results obtained from individual models indicate that different modeling techniques may have different advantages for predicting the AR antagonists. Although the performances of our individual models are comparable, it is difficult to decide which model is the best one and which model should be chosen as a predictor for new compounds. Considering the fitting ability on training set, the *k*NNb model is the best. As for the performance on the prediction set, especially the ability to recognize active compounds, the lazy and ADTree models are better. Thus it seems reasonable that the consensus approach can provide a better predictive ability than the individual models. The availability of several possible models, equally reliable for response prediction, highlights the need for methods able to preserve both model quality and diversity for model comparison.^{51,54–56}

In this work, a simple consensus model was developed that integrated all the three individual models (*k*NNb, lazy, and ADTree). The strategy of majority voting (i.e., 2 actives, 1 inactive = active)⁵⁶ was used to give predictions for all compounds. The consensus model gives the best and most satisfactory results for both the training and the external prediction sets. In the ROC plots of Figure 4 and Figure 5, the consensus curves are on the top of all other models with the largest area of 0.777 for the training set and 0.794 for the prediction set. The consensus model also gives the highest

sensitivity of 72.6% and specificity of 82.7% for the training set among all these models and even better results on the external prediction set (Table 4). The high sensitivity of 76.4% indicates that the consensus analysis can recognize more antagonists from a large number of chemicals. Furthermore, the similar values of the overall concordance for the training (78.6%) and the prediction (79.8%) sets indicate the comparable internal and external predictive abilities of the consensus analysis.

3.2.6. Comparison with Literature Results. The models developed in this work and the MultiCASE model from the literature¹³ were compared (Table 4). In Figure 4 of the ROC plots of all the models for corresponding training set, the consensus curve is on the top with the highest area of 0.777. The MultiCASE model and the lazy prediction performed similarly with areas of 0.716 and 0.719, respectively. The curves from different models are not far from each other, which means that their internal predictive abilities are similar. But in the ROC plots for prediction set (Figure 5), it is obvious that the MultiCASE model curve is quite different, very far from the others. Even though the concordance of this model is as high as 91.7%, the area under the curve is just 0.774, lower than the *k*NNa model (0.796) based on the same training 1 data set. The low sensitivity of MultiCASE model penalizes the quality of the model. On the contrary, other models, based on training 2, perform comparably with similar areas under corresponding curves, especially for the higher external sensitivity, which we have maximized in this work.

3.2.7. Predictions on Test 3. We collected 89 new compounds from the literature^{14,17–19} with heterogeneous structures, to further check our models on a blind external set, completely unbiased by the chemical structure influence present in the splitting by structural similarity. The distributions of these compounds in training 2 are shown in the PCA plot (Figure 7), where it is evident that these collected samples are well distributed into the descriptor space of

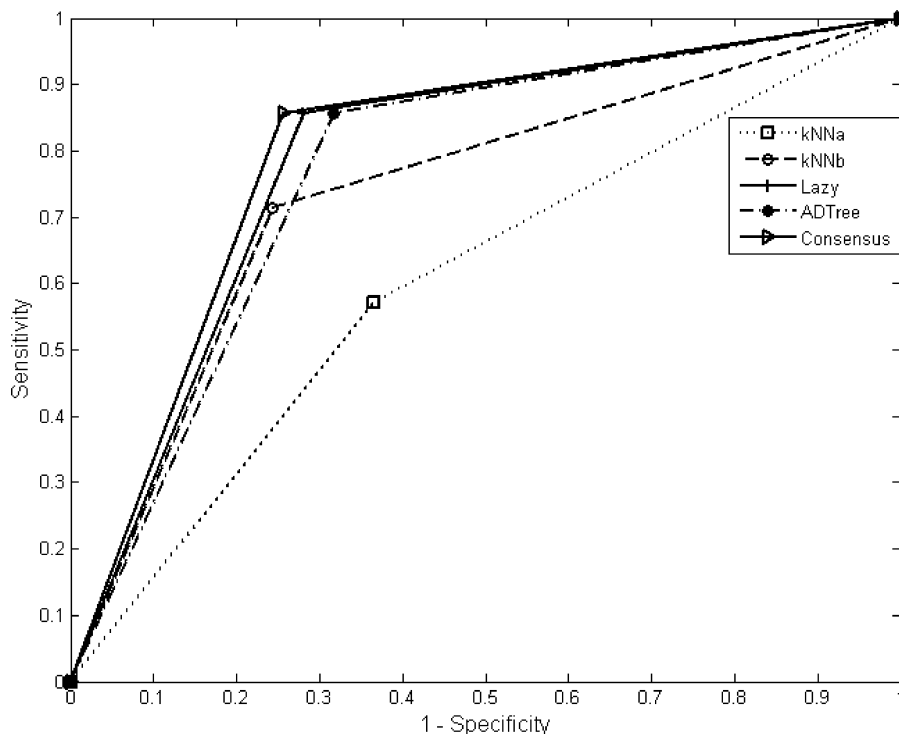


Figure 8. ROC plot on test 3 from different models based on training 2.

training 2. So the predictions for these new compounds should be reliable.

The classification results are also summarized in Table 4. The *k*NNa model, based on training 1, has a much lower performance on test 3 than those of the corresponding prediction set (test 1). The obvious difference between these two prediction sets (test 1 and 3) indicates that the *k*NNa model based on training 1 was not properly evaluated by test 1. This is proof of the poor representativity of test 1, and it is consistent with the statement that the poor external predictive power of the QSAR models could be due to the incorrect usage of external validation.²⁴

On the contrary, all the models based on the new set training 2 perform much better than *k*NNa, especially the consensus and lazy models, which are on the top of the ROC plot for test 3 (Figure 8); the areas under the corresponding curves are as high as 0.801 and 0.788, respectively. In this plot, it is evident that the *k*NNa curve is at the bottom of all other models with the lowest curve area of 0.603. Among the collected data, seven of them are experimentally active. The lazy, ADTree, and consensus approaches predicted six of them correctly with a sensitivity as high as 85.7%. Overall, the consensus approach gave a specificity of 74.4% and a concordance of 75.3%. Comparing the model performance on tests 2 and 3, we can see that the models based on training 2 perform quite similar, which means these models were validated properly by test 2. The predicted assignments by using the models based on training 2 are listed in Table 1.

Two compounds with CAS 521-18-6 and 68047-06-3 in test 3 had discordant experimental activities. They were anticipated inactive in Araki's paper¹⁷ but active (AR binders) according to Fang's work.¹⁴ Two of our models (*k*NNb and ADTree) predict compound 521-18-6 as inactive, and one model predicts it as active. Though the consensus results for this compound are consistent with ref 17, more experiments could be necessary to verify its real

activity. As for compound 68047-06-3, all our models predict it as active, consistent with Fang's work,¹⁴ which means that it could be reasonably an AR antagonist.

3.2.8. Distinction of AR Binders between Agonists and Antagonists. As stated above, 95 AR binders reported by Fang et al.¹⁴ could be AR agonists or antagonists. Mainly due to the limited knowledge about the receptor conformation upon antagonist binding, it remains unclear what kind of ligand–receptor interaction determines the agonist or antagonist activity of the ligand.⁵⁷ To identify more antagonists is also helpful to the analysis of the relationship between their structures and corresponding activities, so as to understand the mechanism of their interaction with androgen receptors.

The distributions of these AR binders are shown in Figure 7, where they are all into the chemical space of training 2. After thorough validation, all the models based on training 2 were used to distinguish these AR binders. The predicted results are summarized in Table 2, where agonists are represented as “G” and antagonists are represented as “T”. Among these binders, 31 of them are predicted as antagonists, and 15 are predicted as agonists by all the three models (*k*NNb, lazy, and ADTree). The testosterone (CAS 58-22-0), which is reported predominantly agonist in ref 14, was among the 15 correctly predicted samples. If the consensus approach is used, 56.8% binders (54 compounds) in this data set are predicted as antagonists, and 43.2% binders (41 compounds) are predicted as agonists.

3.2.9. Virtual Screening of AR Antagonists in Big Data Sets. In the combined data set of MET1 and EDKB, the two biggest data sets of experimental data for estrogen receptor binders, there are 708 compounds. Defined by the leverage approach (with cut off $h \leq 0.0726$), 686 of them are into the applicability domain of our models, including 267 ER binders and 419 nonbinders. By screening this data set using

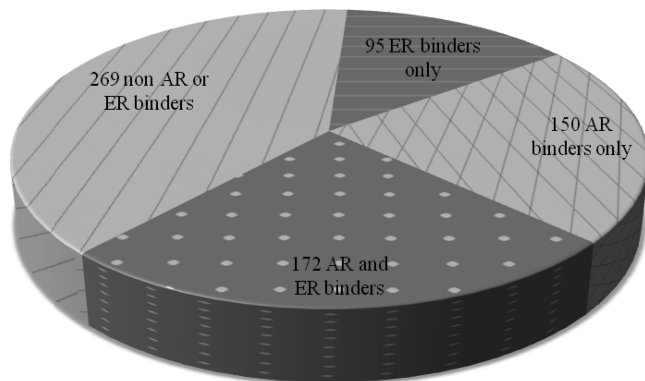


Figure 9. The pie chart of virtual screening results for AR binding on ER binders (as AR binders).

our models (*k*NNb, lazy, and ADTree), we can find out the chemicals that can bind to both androgen and estrogen receptors.

The obtained results are shown in Figure 9. According to our predictions, 127 ER binders were predicted actives as AR binders by all our three models (*k*NNb, lazy, and ADTree), and the consensus approach suggested 172 compounds that can bind to both AR and ER. These compounds may affect the health of humans and wildlife through complex mechanisms, and more research is needed to understand their action with AR and ER so as to forbid or reduce their use. Additionally 150 nonER binders were predicted as AR binders, thus these chemicals that are not considered as endocrine disruptors, based on ER binding, could be EDCs for their possibility to bind to AR. This result increases the number of potential EDCs that are SVHC in REACH. Other 95 ER binders and 269 nonER binders were predicted inactive to AR.

Within our in-house environmental chemicals database, 2487 compounds were found to be within the applicability domain of the proposed classifiers. Approximately 9.7% (241) of chemicals were predicted active for AR antagonism by all three models, which is a percentage similar to the result found in the literature.¹³ If the consensus approach was used, a total of 488 chemicals, about 20% of the screened environmental compounds, were predicted as AR antagonists.

Our screening results suggest that all the *in silico* active chemicals, without experimental activity values, have the maximum possibility to bind to an androgen receptor, so the priority should be given to them to experimentally test the binding activities with androgen receptors.

4. CONCLUSIONS

In this work, we investigated the quantitative structure–activity relationship (QSAR) of a big AR antagonist data set. The strictly externally validated QSAR models were used to distinguish AR binders as agonists and antagonists. We successfully increase the external sensitivity of classification models from 57.1% (as in literature) to 76.4%, and the predictions for additional heterogeneous compounds are even as high as 85.7%. Thus the models proposed here were properly externally validated. We have analyzed the structural information included in the modeling descriptors, and the molecular shape and the electronegative properties are recognized as the most important two factors corresponding to the activity.

We have also applied our predictive models to virtually screen two databases for potential androgen receptor (AR) antagonists. The screening results suggest that about 172 estrogen receptor binders can also bind to androgen receptor, thus they may affect a human's endocrine system with highly complex mechanisms, while 150 compounds, even if not estrogen receptor binders, are predicted as endocrine-disrupting chemicals (EDCs) as AR binders. About 20% (488) of chemicals in a big data set of environmental chemicals are *in silico* predicted as new AR antagonists. The results indicate that our models could be useful as supporting tools for the identification, prioritization, and regulation of new potential EDCs among already existing chemicals. We strongly believe that this QSAR approach is particularly useful not only for identifying substances of very high concern (SVHC) as EDCs, but also for the environmentally benign design of safer replacement solutions for recognized AR disruptors. No method other than QSAR is applicable to chemical design and to detect *a priori*, from the drawn structures, the potential androgen behavior of completely new compounds.

ACKNOWLEDGMENT

Financial supports from the PRIN 07 MIUR Program (2007R57KT7) and the China Scholarship Council (CSC) Postgraduate Study Abroad Program (2008618039). We thank Ester Papa, Barun Bhatarai, and Lili Xi for the useful discussions and the English language revision.

Supporting Information Available: The studied chemicals taken from ref 13 together with the corresponding experimental and predicted classes are listed in Table SII. The combined data of METI and EDKB databases and corresponding predicted classes are summarized in Table SI2. The correlation matrix of the selected descriptors based on training 1 and 2 are shown in Tables SI3 and SI4, respectively. This information is available free of charge via the Internet at <http://pubs.acs.org>

REFERENCES AND NOTES

- (1) Colborn, T. Environmental estrogens: health implications for humans and wildlife. *Environ. Health Perspect.* **1995**, *103*, 135–136.
- (2) Tamura, H.; Ishimoto, Y.; Fujikawa, T.; Aoyama, H.; Yoshikawa, H.; Akamatsu, M. Structural basis for androgen receptor agonists and antagonists: Interaction of SPEED 98-listed chemicals and related compounds with the androgen receptor based on an *in vitro* reporter gene assay and 3D-QSAR. *Bioorg. Med. Chem.* **2006**, *14*, 7160–7174.
- (3) Weintraub, B. D. In *Molecular Endocrinology: Basic Concepts and Clinical Correlations*, 1st ed.; Raven Press: New York, 1995; pp 195–215.
- (4) *Drug-Receptor Interactions*, MERCK company; <http://www.merck.com/mmpe/sec20/ch304/ch304b.html>. Accessed September 7, 2009.
- (5) Tsai, M.; O'Malley, B. W. Molecular Mechanisms of Action of Steroid/Thyroid Receptor Superfamily Members. *Annu. Rev. Biochem.* **1994**, *63*, 451–486.
- (6) Zhou, Z. X.; Wong, C. I.; Sar, M.; Wilson, E. M. The androgen receptor: an overview. *Recent Prog. Horm. Res.* **1994**, *49*, 249–274.
- (7) Araki, N.; Ohno, K.; Nakai, M.; Takeyoshi, M.; Iida, M. Screening for androgen receptor activities in 253 industrial chemicals by *in vitro* reporter gene assays using AR-EcoScreen™ cells. *Toxicol. in Vitro* **2005**, *19*, 831–842.
- (8) Liu, H. X.; Papa, E.; Gramatica, P. QSAR Prediction of Estrogen Activity for a Large Set of Diverse Chemicals under the Guidance of OECD Principles. *Chem. Res. Toxicol.* **2006**, *19*, 1540–1548.
- (9) Liu, H. X.; Papa, E.; Walker, J. D.; Gramatica, P. *In silico* screening of estrogen-like chemicals based on different nonlinear classification models. *J. Mol. Graphics Modell.* **2007**, *26*, 135–144.

- (10) Roncaglioni, A.; Piclin, N.; Pintore, M.; Benfenati, E. Binary classification models for endocrine disrupter effects mediated through the estrogen receptor. *SAR QSAR Environ. Res.* **2008**, *19*, 697–733.
- (11) Liu, H. X.; Yao, X. J.; Gramatica, P. The Applications of Machine Learning Algorithms in the Modeling of Estrogen-Like Chemicals. *Comb. Chem. High Throughput Screening* **2009**, *12*, 490–496.
- (12) Li, J. Z.; Gramatica, P. The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. *Mol. Diversity* **2009**. In press, DOI: 10.1007/s11030-009-9220-2.
- (13) Vinggaard, A. M.; Niemelä, J.; Wedebye, E. B.; Jensen, G. E. Screening of 397 Chemicals and Development of a Quantitative Structure-Activity Relationship Model for Androgen Receptor Antagonism. *Chem. Res. Toxicol.* **2008**, *21*, 813–823.
- (14) Fang, H.; Tong, W. D.; Branham, W. S.; Moland, C. L.; Dial, S. L.; Hong, H. X.; Xie, Q.; Perkins, R.; Owens, W.; Sheehan, D. M. Study of 202 Natural, Synthetic, and Environmental Chemicals for Binding to the Androgen Receptor. *Chem. Res. Toxicol.* **2003**, *16*, 1338–1358.
- (15) REACH; European Chemicals Agency: Helsinki, Finland; <http://europa.eu.int/comm/environment/chemicals/reach.htm>. Accessed June 10, 2009.
- (16) Tong, W.; Welsh, W. J.; Shi, L.; Fang, H.; Perkins, R. Structure-activity relationship approaches and applications. *Environ. Toxicol. Chem.* **2003**, *22*, 1680–1695.
- (17) Araki, N.; Ohno, K.; Takeyoshi, M.; Iida, M. Evaluation of a rapid in vitro androgen receptor transcriptional activation assay using AR-EcoScreen(TM) cells. *Toxicol. in Vitro* **2005**, *19*, 335–352.
- (18) Kojima, H.; Katsura, E.; Takeuchi, S.; Niiyama, K.; Kobayashi, K. Screening for Estrogen and Androgen Receptor Activities in 200 Pesticides by In Vitro Reporter Gene Assays Using Chinese Hamster Ovary Cells. *Environ. Health Perspect.* **2004**, *112*, 524–531.
- (19) Schreurs, R. H. M. M.; Sonneveld, E.; van der Saag, P. T.; van der Burg, B.; Seinen, W. Examination of the in vitro (anti)estrogenic, (anti)androgenic and (anti)dioxin-like activities of tetralin, Indane and isochroman derivatives using receptor-specific bioassays. *Toxicol. Lett.* **2005**, *156*, 261–275.
- (20) Vinggaard, A. M.; Bonefeld Joergensen, E. C.; Larsen, J. C. Rapid and Sensitive Reporter Gene Assays for Detection of Antiandrogenic and Estrogenic Effects of Environmental Chemicals. *Toxicol. Appl. Pharmacol.* **1999**, *155*, 150–160.
- (21) Körner, W.; Vinggaard, A. M.; Térouanne, B.; Ma, R. S.; Wieloch, C.; Schlumpf, M.; Sultan, C.; Soto, A. M. Interlaboratory Comparison of Four in Vitro Assays for Assessing Androgenic and Antiandrogenic Activity of Environmental Chemicals. *Environ. Health Perspect.* **2004**, *112*, 695–702.
- (22) Klopman, G. MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176–184.
- (23) Zhang, Q. Y.; Hughes-Oliver, J. M.; Ng, R. T. A Model-Based Ensembling Approach for Developing QSARs. *J. Chem. Inf. Model.* **2009**, *49*, 1857–1865.
- (24) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.
- (25) Kachigan, S. K. *Multivariate Statistical Analysis: A Conceptual Introduction*. Radius Press: New York, 1991.
- (26) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update *SIGKDD Explorations* **2009**, *11* (1).
- (27) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON—Software for the calculation of molecular descriptors*, version 5.5 for Windows; Taletto S.R.L.: Milan, Italy, 2007.
- (28) METI, ministry of economy trade and industry, Japan. Current status of testing methods development for endocrine disrupters. 6th meeting of the task force on endocrine disrupters testing and assessment (EDTA), Tokyo, Japan June 24–25, 2002; <http://www.meti.go.jp/interface/honsho/Search/English/search?query=gEndocappendix1e&whence=0&max=20&result=normal&sort=score&idxname=meti>. Accessed September 10, 2008.
- (29) FDA EDKB database; Food and Drug Administration (FDA): Silver Spring, MD; <http://edkb.fda.gov/databasedoor.html>. Accessed March, 2009.
- (30) PubChem database; Chemical Abstracts Service (CAS): Columbus, OH; <http://www.ncbi.nlm.nih.gov/pccompound>. Accessed October 17, 2009.
- (31) Zupan, J.; Novic, M.; Ruisánchez, I. Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 1–23.
- (32) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed.* **1993**, *32*, 503–527.
- (33) HYPERCHEM, release 7.03 for Windows; Autodesk, Inc.: Sausalito, CA, 2002.
- (34) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inform. Comput. Sci.* **1989**, *29*, 225–227.
- (35) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.-Act. Relat.* **1997**, *16*, 309–314.
- (36) Todeschini, R.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of the WHIM descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113–119.
- (37) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inform. Comput. Sci.* **2002**, *42*, 682–692.
- (38) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. Wiley-VCH: Weinheim, Germany, 2009.
- (39) Vapnik, V. N. *Statistical learning theory*. Wiley Interscience: New York, 1998.
- (40) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422.
- (41) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*. Wiley Interscience: New York, 1986.
- (42) Zheng, W. F.; Tropsha, A. Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inform. Comput. Sci.* **2000**, *40*, 185–194.
- (43) Freund, Y.; Mason, L. In The alternating decision tree learning algorithm, Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia, June 27–30 1999; pp 124–133.
- (44) Wang, H.; Klinginsmith, J.; Dong, X.; Lee, A. C.; Guha, R.; Wu, Y.; Crippen, G. M.; Wild, D. J. Chemical Data Mining of the NCI Human Tumor Cell Line Database. *J. Chem. Inf. Model.* **2007**, *47*, 2063–2076.
- (45) Liu, Y. Feature extraction and dimensionality reduction for mass spectrometry data. *Comput. Biol. Med.* **2009**, *39*, 818–823.
- (46) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
- (47) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
- (48) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
- (49) *SCAN - Software for Chemometric Analysis*, release 1.1 for Windows; Minitab: State College, PA, 1995.
- (50) 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.
- (51) Gramatica, P.; Pilutti, P.; Papa, E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training-Test Sets and Consensus Modeling. *J. Chem. Inform. Comput. Sci.* **2004**, *44*, 1794–1802.
- (52) Roy, P. P.; Leonard, J. T.; Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 31–42.
- (53) Pearlman, R.; Smith, K. Novel Software Tools for Chemical Diversity, *3D QSAR in Drug Design*; Springer: Kluwer/ESCOM, Dordrecht, Netherlands, 1998; pp 339–353.
- (54) Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discriminant Analysis. *J. Chem. Inf. Model.* **2006**, *46*, 2110–2124.
- (55) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2005**, *46*, 277–288.
- (56) Sutherland, J. J.; Weaver, D. F. Development of Quantitative Structure-Activity Relationships and Classification Models for Anticonvulsant Activity of Hydantoin Analogues. *J. Chem. Inform. Comput. Sci.* **2003**, *43*, 1028–1036.
- (57) Gao, W.; Bohl, C. E.; Dalton, J. T. Chemistry and Structural Biology of Androgen Receptor. *Chem. Rev.* **2005**, *105*, 3352–3370.