# Machine Learning Consensus To Predict the Binding to the Androgen Receptor within the CoMPARA Project

Francesca Grisoni,*[ID] Viviana Consonni,[ID] and Davide Ballabio

Milano Chemometrics & QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, piazza della Scienza 1, IT-20126 Milano, Italy

ABSTRACT: The nuclear androgen receptor (AR) is one of the most relevant biological targets of Endocrine Disrupting Chemicals (EDCs), which produce adverse effects by interfering with hormonal regulation and endocrine system functioning. This paper describes novel in silico models to identify organic AR modulators in the context of the Collaborative Modeling Project of Androgen Receptor Activity (CoMPARA), coordinated by the National Center of Computational Toxicology (U.S. Environmental Protection Agency). The collaborative project involved 35 international research groups to prioritize the experimental tests of approximatively 40k compounds, based on the predictions provided by each participant. In this paper, we describe our machine learning approach to predict the binding to AR, which is based on a consensus of a multivariate Bernoulli Naive Bayes, a Random Forest, and N-Nearest Neighbor classification models. The approach was developed in compliance with the Organization of Economic Cooperation and Development (OECD) principles, trained on 1687 ToxCast molecules classified according to 11 in vitro assays, and further validated on a set of 3,882 external compounds. The models provided robust and reliable predictions and were used to gather novel data-driven insights on the structural features related to AR binding, agonism, and antagonism.
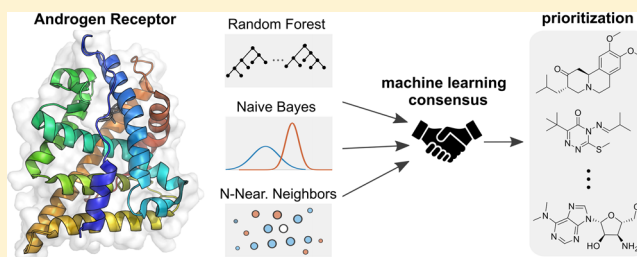
INTRODUCTION

The nuclear androgen receptor (AR), whose gene is located on the X chromosome, is expressed in a wide range of tissues and plays a fundamental biological role in bone, muscle, prostate, adipose tissue, and the reproductive, cardiovascular, immune, neural, and hemopoietic systems.[1,2] AR is one of the target receptors of the so-called Endocrine Disrupting Chemicals (EDCs), exogenous compounds able to disturb hormonal regulation and the endocrine system functioning, thereby producing adverse effects in humans and wildlife.[3,4] EDCs can interact directly with a given nuclear receptor and perturbate or modulate downstream gene expression,[5] but they can also have direct effects on genes and epigenetic impact.[6,7] Disruption of AR-mediated processes can cause irreversible consequences in human health.[6] For example, some chemicals, like pesticides (e.g., DDT), disrupt male reproductive development and function by inhibiting androgen-receptor-mediated events.[3]

Recently, machine learning (ML) and computer-aided techniques have shown to be useful for modeling nuclear receptor modulation of chemicals at different levels, such as drug discovery and design[8−11] and testing prioritization campaigns.[12−14] In the context of the prioritization of hazardous chemicals for experimental testing, in 2016, the National Center of Computational Toxicology (NCCT) of the U.S. Environmental Protection Agency (U.S. EPA) has launched a collaborative project (CoMPARA: Collaborative Modeling Project of Androgen Receptor Activity) to develop in silico approaches to identify man-made chemicals that modulate AR (binding, agonism, and antagonism). The project was a large-scale collaboration among 35 international research groups, following the steps of the predecessor project CERAPP[12] (Collaborative Estrogen Receptor Activity Prediction Project), which targeted the Estrogen Receptor.

The aim of the CoMPARA consortium was to identify, within a list of several thousands of man-made chemicals, those with endocrine disruption potential, to be tested with priority in vitro. The U.S. EPA provided the consortium member groups with a common calibration set of 1689 compounds for model development. Successively, the developed models were utilized to predict the modulation of AR of a library of approximatively 40K chemicals for further testing in the context of ToxCast, Tox21, and other U.S. EPA projects. In particular, the individual groups' predictions were aggregated by the NCCT researchers to identify the compounds that are more frequently predicted as AR modulators.

In this work, we present in silico AR binding models developed in the context of CoMPARA by the Milano Chemometrics and QSAR Research group of the University of Milano-Bicocca. The models are based on the application of machine learning techniques of different nature, which are then integrated within a consensus strategy, to increase the prediction reliability and accuracy. The models were developed in compliance with the OECD principles for the validity of

Quantitative Structure—Activity (QSAR) models,[15] to maximize their usefulness for regulatory applications. While QSAR models targeting AR are usually developed on approximately 20−400 compounds, to the best of our knowledge, the CoMPARA data set is among the biggest and more diverse AR sets available for in silico modeling. Additionally, the data set is based on the results of 11 in vitro assays, thus making the experimental response, and the corresponding models, less prone to generate false positives and false negatives. Finally, the results of our study validate the application of machine learning to the prediction of Androgen Receptor binding and show how these methods provide some data-driven insights on the structural features related to AR modulation.

The paper is structured as follows: after introducing materials and methods, the individual models and their consensus are described. The models are then validated on an external set of approximatively 4000 molecules and interpreted to gain data-driven insights into the structural features relevant for AR binding.

2. MATERIALS AND METHODS

2.1. Data. *CoMPARA Calibration Set.* A QSAR-ready data set containing information and experimental data on AR binding, agonism, and antagonism for 1689 chemicals was provided by the U.S. EPA. It was extracted from Tox21/ToxCast[16,17] programs. AR modulation (binding, agonism, and antagonism) was determined in vitro by high-throughput screening on the basis of 11 in vitro assays capturing the modulation of AR signaling pathway.[18] The participants were asked to provide models for at least one molecular property. We chose AR binding (binding vs nonbinding molecules) due to its best balance between the two classes (13.3% of binders, compared to 10% and 2% of antagonists and agonists, respectively) and the largest number of molecules with annotated values. Two molecules were removed due to issues in the structure annotation as detected by Dragon 7[19] (molecule IDs: DTXSID9035175, DTXSID7032630), thus 1687 molecules were utilized for model development. Data were randomly split into a training (75%, 1265 molecules) and test (25%, 422 molecules) set by stratified sampling, that is, by preserving the proportion between the classes in each set. The training set was utilized for model calibration and internal cross-validation, while the test set was utilized only at the final stage to validate the chosen models.

CoMPARA External Validation Set. The final models were further validated on a set of 3882 compounds, provided at a later stage and used by the U.S. EPA to score the models of the CoMPARA consortium members. This validation set was built by the U.S. EPA from ScrubChem,[20] by considering target, modality, hit quality, and ratio of agreement among the annotated values.[21] The percentage of binders of this set is similar to that of the calibration set (12% of binders over the total). In this work, the external validation set served only to provide an additional assessment of the models' predictivity toward unknown compounds, and, thus, it was not used for model calibration.

2.2. Modeling and Validation. *Molecular Descriptors.* Molecular descriptors[22] were utilized to numerically capture diverse types of structural features of molecules. Two types of molecular descriptors were computed:

1. Extended Connectivity binary fingerprints[23] (ECFPs), also known as "circular fingerprints", which consider the presence of branched molecular substructures and encode it into a binary vector of predefined length through a hashing algorithm. Fingerprints of 1024 bits, with 2 bits per fragment and a radius comprised between 0 and 2 bonds were generated with Dragon 7,[19] with default settings (Count fragments = True, Atom Options: [Atom type, Aromaticity, Connectivity total, Charge, Bond order]).

2. Chemically Advanced Template Search[24] (CATS) descriptors, which capture the distribution of pharmacophores (hydrogen-bond donor and acceptor, positively and negatively charged, and lipophilic atoms) at an increasing topological distances (from 0 to 9 bonds, leading to a total of 150 descriptors). CATS have been developed to capture pharmacophore patterns of chemicals regardless of their molecular scaffolds and, thus, represent a "fuzzier" description of the molecular structure than ECFPs.[24] Also CATS were computed using Dragon 7.[19]

Machine Learning Approaches. We used three different classification approaches to distinguish AR binding and nonbinding molecules, as described below.

1. N-Nearest Neighbors (N3)[25] is a local modeling method similar to the well-established *k*-Nearest Neighbor (*k*-NN).[26] N3 utilizes all training molecules to determine the class of the target molecule and, through an optimized α exponent, tunes their contribution as exponentially decreasing with decreasing their similarity to the target. For any *i*-th target molecule to be predicted, the training molecules are ranked according to their structural similarity to the target, and the score (weight) for any *g*-th class (w_{ig}) is computed as

$$w_{ig} = \frac{\sum_{k=1,k\neq i}^{n} \frac{s_{ik}}{r_{ik}^{\alpha}} \cdot \delta_{kg}}{\sum_{k} \delta_{kg}} \qquad \delta_{kg}$$

$$= \begin{cases} 1 \text{ if } C_k = g \wedge \dfrac{s_{ik}}{r_{ik}^{\alpha}} > \varepsilon \\ 0 \text{ otherwise} \end{cases} \qquad (1)$$

where *k* runs over the training molecules (different from *i*), s_{ik} is the similarity between *i* and *k* (with $s_{ik} \in [0,1]$), and r_{ik} is the (tied) rank of the *k*-th molecule. The Dirac's delta (δ_{kg}) is equal to 1 if *k* belongs to the *g*-th class ($C_k = g$) and if the ratio s_{ik}/r_{ik}^{α} is larger than a fixed ε ($\varepsilon = 10^{-7}$ in this work). The weights are then normalized between 0 and 1 as follows:

$$w'_{ig} = \frac{w_{ig}}{\sum_{g} w_{ig}} \qquad (2)$$

Finally, the *i*-th target is assigned the class with the highest scaled weight (w'_{ig}). The value of α was optimized in 5-fold cross-validation from $\alpha = 0.1$ to $\alpha = 1.5$, with a step of 0.05, and chosen to maximize the cross-validation NER (see eq 9).

2. Random Forest (RF)[27] is a meta-classifier based on an ensemble of several classification trees. Each classification tree is trained on various subsamples of the calibration set, which are built by bootstrapping (*i.e.*, the data set size is preserved by sampling molecules with replacement). The final class is predicted as a majority vote on the trees of the forest. In this work, RF was applied utilizing 100 trees.

Table 1. Summary of the Performance of the Chosen Models in Fitting, 5-Fold Cross-Validation, and on the Test Set[a]

	fitting (1265 chemicals)				cross-validation (1265 chemicals)				test set (422 chemicals)			
ID	Sn	Sp	NER	np%	Sn	Sp	NER	np%	Sn	Sp	NER	np%
N3	0.83	0.79	0.81	0	0.78	0.81	0.79	0	0.85	0.75	0.80	3
RF	1.00	0.83	0.91	0	0.74	0.75	0.75	0	0.70	0.70	0.70	1
NB	0.86	0.71	0.78	0	0.82	0.70	0.76	0	0.65	0.80	0.72	3
Consensus Strict	0.91	1.00	0.96	29	0.81	0.88	0.84	30	0.92	0.85	0.83	37
Consensus Loose	0.83	0.85	0.84	0	0.78	0.81	0.81	0	0.76	0.78	0.75	3

[a]The model parameters were optimized in 5-fold cross-validation (N3: $\alpha = 1.00$; RF: misclassification cost of binder molecules, $C = 0.96$). Sensitivity (Sn), specificity (Sp), nonerror rate (NER), and percentage of compounds not predicted (np%) are reported.

Trees were grown by minimizing the nodes' impurity, as measured by the Gini's diversity index (GDI).[28] Let m denote the m-th node of the tree and GDI_m the node's impurity, GDI_m is defined as follows

$$\mathrm{GDI}_m = 1 - \sum_g \left(\frac{n_{mg}}{n_m} \right)^2 \tag{3}$$

where the summation runs over the classes, and n_{mg}/n_m denotes the fraction of molecules in the m-th node belonging to the g-th class. A "pure" node (i.e., a node containing only molecules with the same experimental class) will have GDI = 0; the larger the GDI_m, the less pure the m-th node. Due to the unbalanced partition between the two classes (13.3% of binders and 86.7% of nonbinders), in analogy with a recent study,[29] a misclassification cost for binders (i.e., penalty for generating false negatives) was introduced and optimized in 5-fold cross-validation (varying from 0.90 to 0.98 with a step of 0.02), along with the number of trees (tested values = [50, 100, 500]).

3. Bernoulli Naive Bayes (NB)[30] is a conditional probability approach based on the Bayes' theorem. Given an input i (i.e., the target molecule), described by a vector of p features $\mathbf{x}_i = (x_{i1},...,x_{ip})$, NB calculates the posterior probability for each g-th class, expressed as

$$P(C_i = g | x_{i1}, ..., x_{ip}) = P(C_i = g | \mathbf{x}_i) \tag{4}$$

According to the Bayes' theorem, the posterior probability can be calculated as follows

$$P(C_i = g | \mathbf{x}_i) = \frac{P(g) P(\mathbf{x}_i | C_i = g)}{P(\mathbf{x})} \tag{5}$$

where $P(g)$ is the prior probability, $P(\mathbf{x}_i | C_i = g)$ is the class-conditional probability, and $P(\mathbf{x})$ is the input feature probability. The prior probability of the g-th class, $P(g)$, is the relative occurrence frequency of the g-th class

$$P(g) = \frac{n_g}{n} \tag{6}$$

with n_g being the number of molecules belonging to the g-th class, and n being the total number of molecules. $P(\mathbf{x})$, which is the probability of observing values for the input features, is constant and can, thus, be neglected. The class-conditional probability, $P(\mathbf{x}_i | C_i = g)$, is the probability of observing a given set of input features ($\mathbf{x}_i$) in the molecules of the g-th class, and it can be estimated by several methods.[31] In this work, we used the multivariate Bernoulli NB approach[32] for binary data; the class-conditional probability was computed as

$$P(\mathbf{x}_i | C_i = g) = \prod_{j=1}^{p} P(g | x_j)^{x_{ij}} [1 - P(g | x_j)]^{1 - x_{ij}} \tag{7}$$

where $P(g | x_j)$ is the posterior probability of the j-th feature in the data set, computed as

$$P(g | x_j) = \frac{1 + \sum_{k=1}^{n} \delta_{kg}}{G + n_g} \text{ where } \delta_{kg}$$
$$= \begin{cases} 1 \text{ if } C_k = g \wedge x_{kj} = 1 \\ 0 \text{ otherwise} \end{cases} \tag{8}$$

where k runs over the n training molecules, n_g is the number of molecules belonging to the g-th class, G is the number of classes, and δ_{kg} is equal to 1 if the k-th molecule has the j-th feature and belongs to the g-th class. In other words, $P(g | x_j)$ is the occurrence frequency of the j-th feature in the considered class. The values of 1 and G at the numerator and denominator, respectively, are referred to as Laplacian smoothing, utilized to rule out zero-valued conditional probabilities.[33] Intuitively, given a new molecule represented by a set of binary descriptors, NB calculates the posterior probability for each class (eq 4) based on (i) the class prior probability (the higher, the higher the posterior probability) and (ii) the occurrence frequency of the molecular features in that class (the higher the frequency, the higher the posterior probability). The molecule is then assigned the class with the highest posterior probability based on the observed molecular features.

Statistical Evaluation of the Model Performance. Models were evaluated through three classification indices, namely Sensitivity (Sn), Specificity (Sp), and Non-Error Rate (NER), defined as follows[34]

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{NER} = \frac{\mathrm{Sn} + \mathrm{Sp}}{2} \tag{9}$$

where TP and TN quantify the number of correctly classified positive compounds (binders) and correctly classified negative compounds (nonbinders), respectively, while FP and FN quantify the number of misclassified negative (nonbinders) and positive (binders) compounds, respectively. While Sn and Sp refer to the two classes of interest (fraction of correctly classified binders and nonbinders, respectively), the NER (also known as balanced accuracy or, in the case of binary

classification tasks, as Area Under the ROC Curve[34]) quantifies the overall classification performance. The NER avoids biased evaluations when the classes are unbalanced.[34] Sn, Sp, and NER were calculated in fitting, 5-fold cross-validation (venetian blind resampling), and on the test and external sets.

2.3. Software and Code. Molecular descriptors were computed with Dragon 7.[19] For calculating CATS descriptors and radial fragments, the molecules were standardized using RDKit "Structure Normalizer" node (default settings) of Konstanz Information Miner (KNIME)[36] platform, to ensure the correct identification of functional groups. All the calculations were performed in MATLAB[35] using code written by the authors. Random Forest predictions were built on MATLAB "TreeBagger" function. The Minimum Spanning Tree (MST) was calculated with the MATLAB "*minspan*" function and displayed with Pajek (v 5.05).[37] Data and MATLAB code for applying Naive Bayes, N3, and Random Forest are provided for free on the Milano Chemometrics Web site (http://www.michem.unimib.it/download/data).

RESULTS AND DISCUSSION

Single Models. The machine-learning models were calibrated on the 1265 training set molecules, using different
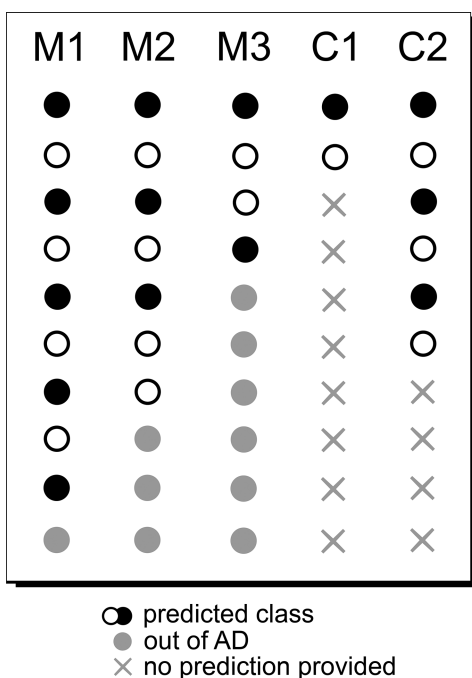
Figure 1. Decision rules of the two consensus approaches, here referred to as Strict and Loose, based on the three single models (M1, M2, and M3). The strict consensus provides a prediction if and only if all three models (M1, M2, and M3) agree. The loose consensus provides a prediction in a majority voting fashion, by choosing the class that is most frequently predicted among the three individual models. Predictions out of the AD are not considered by any of the consensus approaches. Note that M1, M2, and M3 may represent any one of the chosen single models.

combinations of molecular descriptors (ECFPs and CATS) and modeling approaches. The best models resulted to be (i) N-Nearest Neighbors (N3) applied to ECFPs with Jaccard-Tanimoto similarity index and tuning parameter (α, eq 1) equal to 1.00, (ii) Random Forest (RF) applied to CATS descriptors (100 trees, misclassification cost for binder compounds equal to

0.98), and (iii) multivariate Bernoulli Naive Bayes (NB) with Lagrange smoothing applied to ECFPs. The model parameters were chosen with 5-fold cross-validation. The cross-validation protocol also confirmed the model robustness and stability, with a classification performance comparable to that obtained in fitting (Table 1).

The 422 test set molecules served as a second step of model validation. To ensure the reliability of the prediction and rule-out model extrapolations for molecules too different from the training set, each model was equipped with an Applicability Domain (AD) assessment. The AD can be defined as the chemical space where the model predictions can be considered as reliable,[38] and it is an essential aspect to consider when predicting untested compounds according to the OECD (Organization for Economic Cooperation and Development) principles for QSAR.[15] In this study, the AD was defined for each modeling strategy on a case-by-case basis, as follows:

1. N-Nearest Neighbors (N3): a target molecule is considered outside the model applicability domain if both its class scores (eq 2) are lower than the predefined class threshold. The thresholds were defined as the 95th percentile of the class scores of the training molecules ($w^*_{binder} = 0.8789$, $w^*_{nonbinder} = 0.9007$, with $\alpha = 1.00$);

2. Naive Bayes (NB): a target molecule is considered outside the model applicability domain if its average Jaccard-Tanimoto similarity to the nearest 100 training molecules is lower than a defined threshold. This threshold was defined as the 95th percentile of the average Jaccard-Tanimoto similarity of the training chemicals to their 100 closest neighbors ($s^* = 0.1748$);

3. Random Forest (RF): the AD was defined utilizing the bounding-box approach,[38] that is, through a hyper-rectangle delimited by the minimum and maximum value of each descriptor of the training set molecules. If a molecule falls out of such range for at least one descriptor, then it is considered as out of the AD.

All the calibrated models have a comparable performance in cross-validation and on the test set but differ from each other (Table 1). N3 and NB, calibrated on ECFPs, show an opposite performance in modeling AR binding, the former being more suitable in recognizing binder compounds (Sn = 0.85 and Sp = 0.75 on the test set), while the latter having a better performance on nonbinders (Sn = 0.65 and Sp = 0.80 on the test set). The performance differences between the two approaches are due to the different characteristics of the two machine learning methods. N3, in fact, tends to favor the less-numerous classes, due to the normalization over the number of the utilized neighbors belonging to a given class for calculating the score (eq 2). Additionally, N3 was calibrated using the Jaccard-Tanimoto similarity index, which, for two molecules, considers only the presence of common features but neglects the features that are absent in both molecules. On the contrary, NB with the Bernoulli multivariate approach explicitly considers also the absence of features for class assignment (eq 7). RF shows the most balanced performance among the utilized approaches (Sn = Sp = 0.70 on the test set). As typically observed for RF, the model has a good capacity of describing the data when applied in fitting (Sn = 1.00 and Sp = 0.83, Table 1). This is due to the "descriptive" nature of RF individual trees, which are not pruned and, thus, have a good fitting capacity. The lack of overfitting of the final RF model, which is a consensus among the individual bootstrapped trees, is thus shown by the good performance in

Table 2. Summary of the Performance of the Chosen Models after Refitting on the Whole Calibration Set and Their Application to the External Evaluation Set[a]

ID	param	calibration set (1687 chemicals)				external set (3882 chemicals)			
		Sn	Sp	NER	np%	Sn	Sp	NER	np%
N3	$\alpha = 1.00$	0.82	0.80	0.81	0	0.72	0.75	0.74	6
RF	$C = 0.96$	1.00	0.78	0.89	0	0.72	0.75	0.73	3
NB		0.69	0.81	0.75	0	0.65	0.79	0.72	4
Consensus Strict		1.00	0.89	0.94	28	0.80	0.87	0.83	39
Consensus Loose		0.85	0.81	0.83	0	0.71	0.79	0.75	5

[a]Sensitivity (Sn), specificity (Sp), nonerror rate (NER), and percentage of compounds not predicted (np%) are reported.
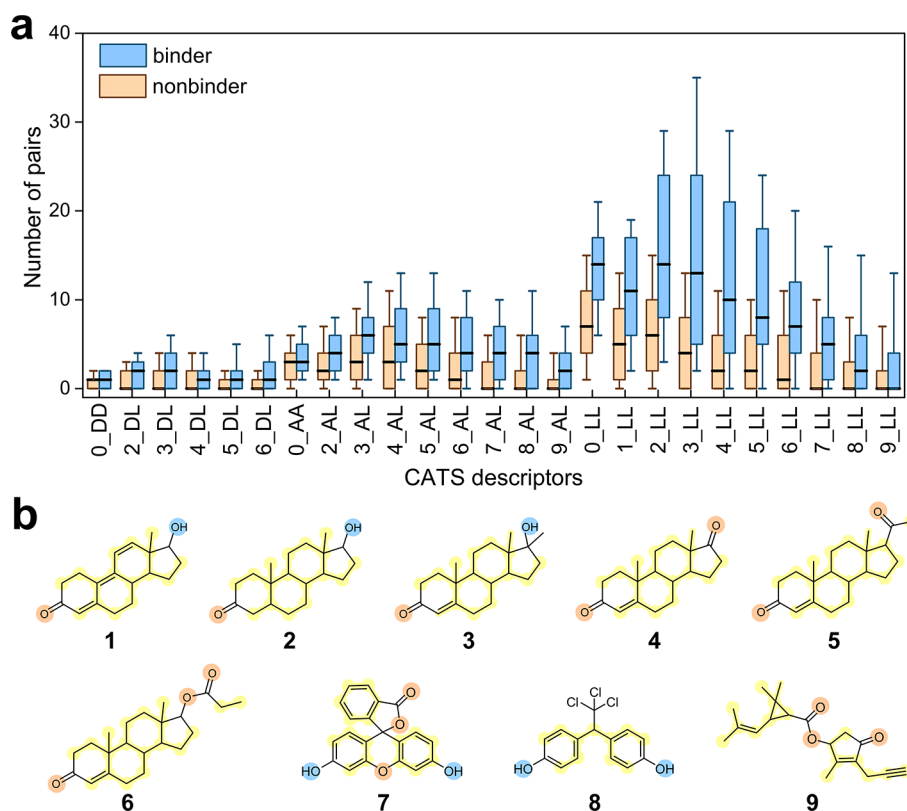
Figure 2. Analysis of CATS descriptors. (a) CATS descriptors most frequently selected by RF (frequency >95%) and significantly different between nonbinder and binder molecules ($p < 0.05$, Mann–Whitney test). Letters indicate the pairs of pharmacophores considered (L: lipophilic, A: acceptor, D: donor), while the number indicates the topological distance between the two pharmacophores (from 1 to 9, 0 represents the occurrence of a pharmacophore itself). Boxplots indicate median (solid line), first and third quartiles (box), 5th and 95th percentiles (whiskers). Colors indicate the experimental class. (b) Pharmacophore patterns of steroidal and nonsteroidal binders present in the data set (orange: acceptor, blue: donor, yellow: lipophilic) labeled in agreement with Dragon CATS scheme. **1** = 17a-trenbolone; **2** = 5alpha-dihydrotestosterone (DHT); **3** = 17-methyltestosterone; **4** = 4-androstene-3,17-dione; **5** = progesterone; **6** = testosterone propionate, **7** = fluorescein; **8** = hydroxychlor; **9** = prallethrin.

cross-validation (Table 1) and its comparability with the statistics obtained on the test set, which was not used to train the model.

Consensus Approaches. To increase the predictivity of the developed models, we applied a consensus approach. Consensus approaches consist of the aggregation of models of different nature and/or with different performance, generally with the aim of improving the overall model predictivity (e.g., refs 39 and 40) and prediction confidence. In this work, two types of consensus modeling were performed:

1. Consensus "Strict", which provides a prediction for a molecule if and only if all three models agree on it (Figure 1). The predicted class will be the one predicted in agreement by all three models. In any other case

(disagreeing predictions or molecule out of at least one AD), no prediction is provided.

2. Consensus "Loose", which operates in the so-called "majority-voting" fashion. In this case, the prediction provided is chosen as the class that is most frequently predicted among the three models. Note that predictions out of the AD are not considered (Figure 1).

The two consensus approaches lead to an improved performance in comparison with the single models on both the training and the test set (Table 1).

The strict consensus, which is the most restrictive, leads to a large improvement in the performance, especially on the test set (up to 0.27 and 0.15 of improvement on Sn and Sp, respectively, Table 1). The restrictive character of the approach causes many compounds to be not predicted (up to 37% on the test set, Table
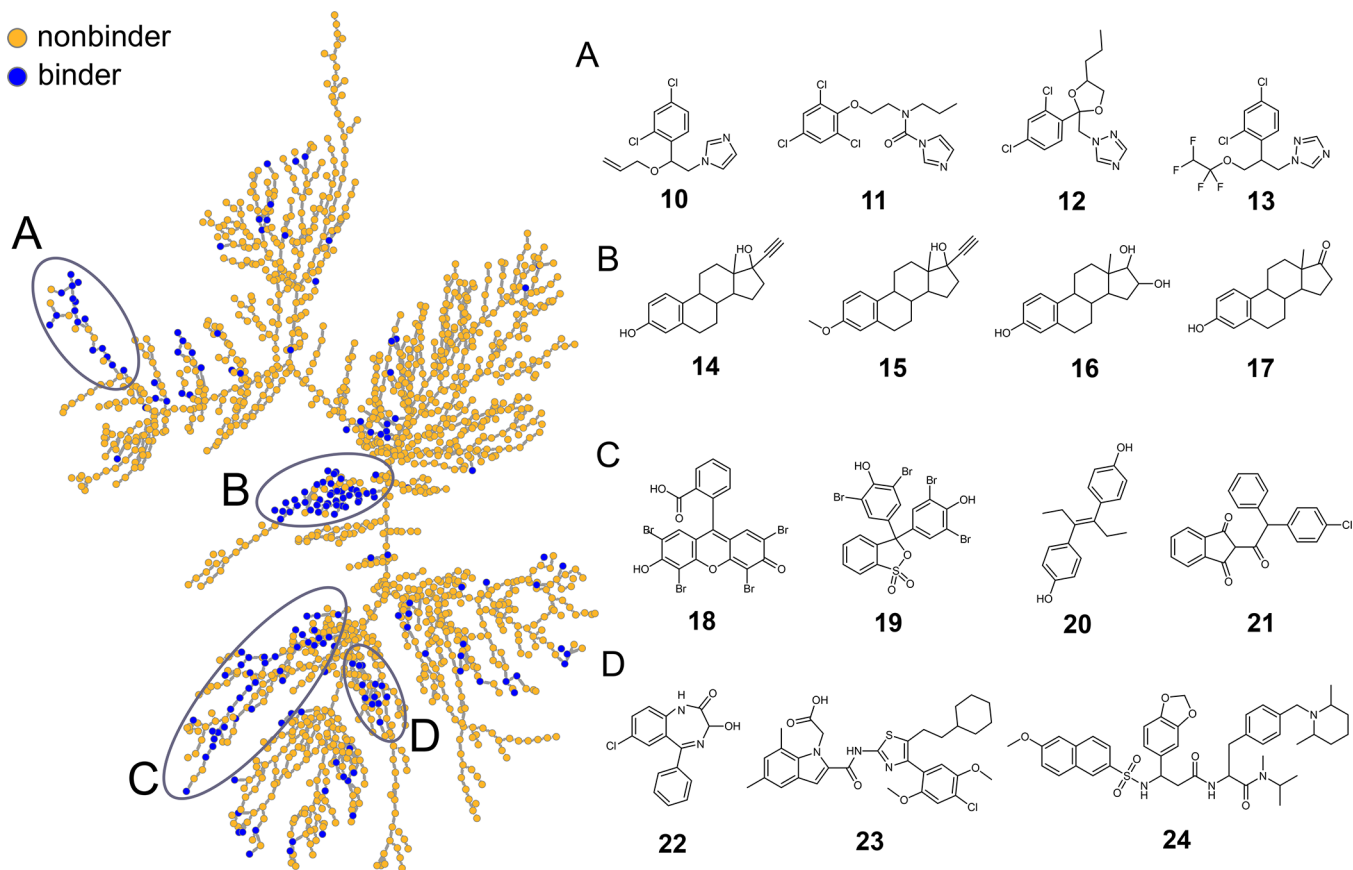
Figure 3. Minimum Spanning Tree performed on the ECFP. Compounds are colored according to their experimental class (orange: nonbinder, blue: binder). The four most numerous clusters of binders are marked with capital letters (from A to D). Some examples of compounds belonging to each cluster are reported on the right (**10** = imazil, **11** = prochloraz, **12** = propiconazole, **13** = tetraconazole, **14** = 17alpha-ethinylestradiol, **15** = mestranol, **16** = estriol, **17** = estrone, **18** = eosin, **19** = bromophenol blue, **20** = diethylstilbestrol, **21** = chlorophacinone, **22** = oxazepam, **23** = SR146131, **24** = SSR240612).

1), due to at least one disagreeing prediction among the single models. This approach is most suited to applications where the reliability of the prediction must be ensured, at the expenses of the number of molecules for which obtaining a prediction is possible.

The loose consensus leads to a more balanced performance than N3 and NB, with a general improvement compared to RF (Table 1). Unlike the strict consensus, the improvement is moderate, but also the number of nonpredicted compounds (up to 3%) is moderate. Thus, such a consensus approach is more suited than the strict consensus to provide predictions for the largest number of compounds possible, when necessary.

The prediction improvement obtained with the consensus approach is due to the aggregation of complementary computational techniques, each grasping partial information on the training data, thereby reaching better predictions than the individual models.[27]
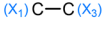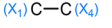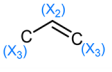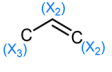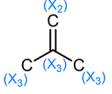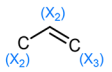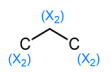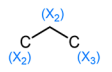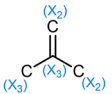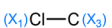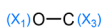
External Validation. In the final stage of the model development process, the models were refitted on the complete calibration set of 1687 molecules with the settings defined previously, and utilized for predicting the 3882 molecules of the external validation set (Table 2). The model AD was redefined with new thresholds determined on the whole calibration set (N3: $w^*_{nonbinder}$ = 0.9042, w^*_{binder} = 0.8645; NB: s^* = 0.1884). The obtained performance is comparable to what was previously observed on the internal test set (Table 1), thus further confirming the model robustness and applicability to novel

compounds. Consensus approaches lead to improved NER values on the external set, with the highest performance (Sn = 0.80, Sp = 0.87, NER = 0.83) obtained by the strict consensus, at the expenses of the number of predicted compounds (only 61% compounds are predicted). The loose consensus leads to the second largest NER (NER = 0.75), with only 5% of compounds not predicted. N3 has a similar nonerror rate (NER = 0.74), with, however, a larger number of nonpredicted compounds (6%) and a smaller Sp (4% smaller). Since high values of Sp are well-suited for the prioritization purposes of the project, the loose consensus was retained as the final chosen approach, despite its only moderate improvement compared to the single models.

Model-Driven Mechanistic Insights. Machine learning algorithms can be leveraged to increase the understandability of the obtained predictions and to gather useful data-driven insights. Thus, in this paragraph, the models are interpreted on a molecular descriptor-basis.

CATS Descriptors. CATS descriptors are intrinsically connected to a structural interpretation, as they encode information on the occurrence of pharmacophore pairs (hydrogen-bond donor [D] and acceptor [A], positively [P] and negatively [N] charged, and lipophilic [L] atoms) at increasing topological distance. The most frequently occurring CATS descriptors in the 100 trees constituting the RF model (frequency >95%, 35 descriptors) were utilized to enhance the model interpretability (Figure 2). A Mann–Whitney test[41] was

Table 3. ECFP-Related Molecular Fragments with Significantly Diverse Occurrence between Nonbinders and Binders ($p < 0.05$, Mann−Whitney Test[a])

ID	Fragments Depiction	SMARTS string	Binders (B) [Mean ± SE]	Nonbinders (nB) [Mean ± SE]	Ratio B/nB [Mean ± SE]
F1	C(X₂)	[C;X2]	9.4 ± 0.3	6.1 ± 0.1	1.54 ± 0.07
F2	C(X₃)	[C;X3]	6.0 ± 0.3	3.4 ± 0.1	1.8 ± 0.1
F3	C(X₄)	[C;X4]	0.93 ± 0.09	0.35 ± 0.03	2.7 ± 0.8
F4	Cl	[Cl;X1]	0.65 ± 0.09	0.34 ± 0.02	2 ± 1
F5	N(X₃)	[N;X3]	0.53 ± 0.06	0.30 ± 0.02	1.8 ± 0.7
F6	O(X₁)	[O;X1]	2.0 ± 0.2	1.60 ± 0.04	1.3 ± 0.3
F7	(X₁)C—C(X₃)	[C;X1][C;X3]	0.31 ± 0.06	0.50 ± 0.03	0.6 ± 1.0
F8	(X₁)C—C(X₄)	[C;X1][C;X4]	0.71 ± 0.08	0.34 ± 0.03	2 ± 1
F9	[depiction]	[C;X2]([C;X3])=[C;X3]	1.0 ± 0.1	0.39 ± 0.02	2.6 ± 0.7
F10	[depiction]	[C;X2]=[C;X2][C;X3]	1.3 ± 0.1	0.76 ± 0.03	1.8 ± 0.4
F11	[depiction]	[C;X2]=[C;X3]([C;X3])[C;X3]	0.26 ± 0.04	0.15 ± 0.01	2 ± 2
F12	[depiction]	[C;X2][C;X2]=[C;X3]	1.7 ± 0.1	1.06 ± 0.03	1.6 ± 0.2
F13	[depiction]	[C;X2][C;X2][C;X2]	0.81 ± 0.2	0.90 ± 0.06	0.9 ± 1.6
F14	[depiction]	[C;X2][C;X2][C;X3]	1.2 ± 0.1	0.39 ± 0.02	3.1 ± 0.8
F15	[depiction]	[C;X2][C;X3](=[C;X2])[C;X3]	0.59 ± 0.07	0.26 ± 0.02	2.3 ± 0.8
F16	(X₁)Cl—C(X₃)	[Cl;X1][C;X3]	0.49 ± 0.06	0.21 ± 0.02	2 ± 1
F17	(X₁)O—C(X₃)	[O;X1][C;X3]	0.6 ± 0.1	0.38 ± 0.02	2 ± 1
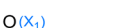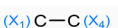
[a]ID, depiction, and SMARTS string are reported, along with the mean value and standard error (SE) calculated on the fragment occurrence frequency (i.e., no. of fragments) in nonbinder and binder molecules, and the ratio between binders and nonbinders. In the fragment depiction, "X" followed by a number represents the atom connectivity (H-depleted).

performed to detect the descriptors within this reduced pool that have significantly different values ($p < 0.05$) in binder molecules compared to nonbinders. The most statistically relevant descriptors are connected to the presence of (i) lipophilic regions (L atoms separated by 1 to 9 bonds), (ii) pairs of acceptor-lipophilic (AL) separated by two to nine bonds, and (iv) donor-lipophilic pairs separated by 2 to 6 bonds. Additionally, binder compounds result in having a higher number of donor ("0_DD"), acceptor ("0_AA"), and lipophilic ("0_LL") atoms (Figure 2a).

The selected descriptors capture well-known pharmacophoric patterns of natural and synthetic steroids, such as testosterone and DHT (Figure 2b, **1**−**6**). Hydrophobic interactions are known to be relevant for the ligand interaction with the hydrophobic pocket of the ligand binding domain (LBD) of AR. Additionally, the acceptor/donor pharmacophoric features are
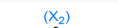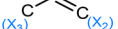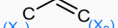
fundamental for positioning and anchoring in the receptor pocket.[42,43] The "fuzzy" character[24] of CATS descriptors allows for also considering nonsteroidal scaffolds, thereby correctly predicting other types of AR-binders, such as the organic dye fluorescein (**7**), hydroxychlor (**8**), and the insecticide prallethrin (**9**), which share similar pharmacophore patterns to steroids. The findings related to our RF model are in agreement with the work of Tamura and coauthors,[44] which identify strong hydrogen bonding and/or electrostatic interaction ability at the position corresponding to the 3-keto group of DHT (**2**) and an H-bond acceptor or donor group at the position corresponding to the 17β-OH of DHT as responsible for the interaction with the ligand binding domain. Additionally, lipophilic groups induce the interaction with the hydrophobic pocket and play a prominent role in increasing AR binding activity.[44]

Table 4. ECFP-Related Molecular Fragments with Significantly Diverse Occurrence between Agonists and Antagonists ($p < 0.05$, Mann−Whitney Test)[a]

ID	Fragments Depiction	SMARTS string	Agonists [Mean ± SE]	Antagonists [Mean ± SE]	Ratio Ag/Ant [Mean ± SE]
F18	C (X_1)	[C;X1]	2.1 ± 0.2	1.7 ± 0.1	1.2 ± 0.1
F19 (F1)	C (X_2)	[C;X2]	10.2 ± 0.5	9.2 ± 0.4	1.10 ± 0.05
F20 (F2)	C (X_3)	[C;X3]	7.6 ± 1.3	5.6 ± 0.3	1.4 ± 0.1
F21 (F3)	C (X_4)	[C;X4]	1.8 ± 0.2	0.7 ± 0.1	2.6 ± 0.4
F22 (F4)	Cl	[Cl;X1]	0.05 ± 0.04	0.8 ± 0.1	0.10 ± 0.02
F23 (F5)	N (X_3)	[N;X3]	0.3 ± 0.1	0.6 ± 0.1	0.4 ± 0.1
F24 (F6)	O (X_1)	[O;X1]	3.6 ± 0.9	1.6 ± 0.1	2.2 ± 0.3
F25	(X_1) C—C (X_4)	[C;X1][C;X4]	1.4 ± 0.2	0.5 ± 0.1	2.6 ± 0.5
F26 (F10)	(X_3)C=C $(X_2)(X_2)$	[C;X2]=[C;X2][C;X3]	0.7 ± 0.2	2.1 ± 0.2	0.30 ± 0.05
F27	(X_2)C=C $(X_2)(X_2)$	[C;X2][C;X2]=[C;X2]	0.4 ± 0.2	0.9 ± 0.1	0.4 ± 0.1
F28 (F12)	(X_2)C=C $(X_3)(X_2)$	[C;X2][C;X2]=[C;X3]	0.5 ± 0.2	1.5 ± 0.1	0.30 ± 0.07
F29 (F14)	(X_2)C—C $(X_2)(X_3)$	[C;X2][C;X2][C;X3]	3.7 ± 0.4	0.5 ± 0.1	7 ± 1
F30	(X_1)O=C (X_3)	[O;X1]=[C;X3]	1.6 ± 0.3	0.6 ± 0.1	2.8 ± 0.4

[a]ID, depiction, and SMARTS string are reported, along with the mean value and standard error (SE) calculated on the fragment occurrence frequency (i.e., number of fragments) in agonist and antagonist molecules. In the fragment depiction, "X" followed by a number represents the atom connectivity (H-depleted). ID in brackets refers to fragments previously identified in Table 3.

ECFPs. To analyze the structural information encoded within ECFPs and grasped by the machine learning models, we performed a Prim's minimum spanning tree (MST),[45] which allows for representing the similarities between the molecular fingerprints in the form of a graph, where the vertices are the molecules and edges are their similarity relationships (Figure 3). MST was calculated utilizing the Jaccard-Tanimoto similarity index, the same used for N3 modeling. In the descriptors' space of ECFPs, AR modulators are successfully clustered, with four well-separated clusters (Figure 3, A−D), thus motivating the satisfactory performance of N3 and NB models. The obtained clusters contain distinct categories of compounds. Cluster A mainly contains compounds characterized by the presence of halogenated benzenes, imidazole/triazole groups, and a hydrophobic chain. Cluster B collects steroidal compounds, such as 1−9 and 17alpha-ethinylestradiol (Figure 3, 14). Clusters C and D are more heterogeneous, and both contain compounds with many halogenated groups and/or atoms with high electrostatic or donor/acceptor potential, features that are relevant for the interaction with AR-LBD.[44]

In analogy with recent studies,[46−48] the fragments generated by the ECFP algorithm were used to enhance the interpretation of the structural features captured by the models; the occurrence of all fragments of the calibration molecules having a radius between 0 and 2 bonds was generated (7773 fragments). Like in the case of CATS, a Mann−Whitney test[41] allowed for the detection of fragments with a significantly different occurrence ($p < 0.05$) in binder molecules compared to nonbinders. Table 3 lists the fragments with significant differences of occurrence in binders and nonbinders and present in at least 10% of the calibration set molecules. Binder compounds have from 1.5 to approximatively three times more secondary, tertiary, and

quaternary carbon atoms (fragments F1 to F3), which contribute to a general increase of the compounds' hydrophobicity and the corresponding interaction with the hydrophobic pocket of AR. Binders contain a higher number of sp^2-hybridized carbon atoms than nonbinders, a feature known to increase the binding affinity to AR-LBD, due to the reduction of molecular flexibility.[44] Additionally, the presence of hydrogen bond donors/acceptors (e.g., F5, F6, and F17) confirms what is already observed on CATS and ECFPs.

Since the mechanism by which AR differentiates between agonist and antagonist ligands is still poorly understood,[49,50] we expanded our structure−activity relationship (SAR) investigation by analyzing ECFP-related fragments on binders labeled as either AR agonist or antagonists (39 and 155 compounds in the calibration set, respectively). Table 4 lists the fragments with significant differences of occurrence in agonists and antagonists ($p < 0.05$) and present in at least 30% of the considered molecules. Agonist molecules have up to twice the number of aliphatic carbons compared to antagonists (F18, F19, F20, and F21, Table 4), especially in the form of linear aliphatic chains (F29: 7 ± 1 times higher on average), as well as a higher number of oxygen atoms (fragment F24), especially in the form of carbonyl groups (fragment F30: 2.8 ± 0.4 times higher) and potentially responsible for the positioning and anchoring in the receptor pocket.[42,43] Antagonists have a lower saturation degree and a higher number of sp^2-hybridized carbon atoms (F26, F27, and F28), as well as a higher number of nitrogen and chlorine atoms (fragments F22 and F23) compared to agonists. These differences provide insights into the structural features related to the conformational changes due to agonist and antagonist binding to AR-LBD[50] and will need further investigation.

CONCLUSION

This work presented a consensus approach based on machine learning models for the prediction of the binding to the androgen receptor (AR). Such models have been developed in the context of the international collaborative project (CoMPARA) promoted by the U.S. EPA, for the prioritization of potential endocrine disrupting chemicals. The models were developed in compliance with OECD principles for QSAR validity and on a reliable data set merging 11 in vitro assays, thus making the experimental response, and the corresponding models, less prone to false positive and false negative generation. The single models (N-Nearest Neighbor [N3], Random Forest [RF], and Naive Bayes [NB]) were developed on approximately 2000 molecules and show different performances, with N3 having the largest sensitivity, NB having the largest specificity, and RF being the most balanced. The consensus models showed an increased prediction reliability compared to the single models, especially on the external validation compounds. The external validation of the consensus approaches on approximately 4000 external molecules confirmed the model predictivity toward unknown compounds, with NER equal to 0.75 and 0.83 for the loose and strict approaches, respectively. The presented models were then utilized to predict the potential for AR binding of approximatively 40,000 untested compounds, to be used by the U.S. EPA for testing prioritization. The models also contribute to the expansion of the chemical space covered by computational models for ToxCast, Tox21, and other U.S. EPA projects. The descriptor-based interpretation of the selected models highlighted some structural features related to AR binding, such as the positioning of pharmacophoric features, the number of sp^2-hybridized carbon atoms, and the presence of hydrogen bond donors/acceptors. Our work shows the potential of machine learning and consensus algorithms for testing prioritization, animal testing reduction, and data-driven mechanistic insights.

AUTHOR INFORMATION

Corresponding Author

*E-mail: francesca.grisoni@unimib.it.

ORCID

Francesca Grisoni: 0000-0001-8552-6615
Viviana Consonni: 0000-0001-6252-9805

Notes

The authors declare no competing financial interest.

Data and Code Availability. The data set with the computed Dragon descriptors and experimental class (binding/nonbinding), along with the MATLAB code for applying Naive Bayes, N3, and Random Forest, is provided for free on Milano Chemometrics Web site (http://www.michem.unimib.it/download/data/ar-binding-compara-project/).

REFERENCES

(1) Rana, K.; Davey, R. A.; Zajac, J. D. Human Androgen Deficiency: Insights Gained from Androgen Receptor Knockout Mouse Models. *Asian J. Androl.* **2014**, *16* (2), 169.

(2) Manolagas, S. C.; O'brien, C. A.; Almeida, M. The Role of Estrogen and Androgen Receptors in Bone Health and Disease. *Nat. Rev. Endocrinol.* **2013**, *9* (12), 699.

(3) Fisher, J. S. Environmental Anti-Androgens and Male Reproductive Health: Focus on Phthalates and Testicular Dysgenesis Syndrome. *Reproduction* **2004**, *127* (3), 305−315.

(4) Henley, D. V.; Korach, K. S. Endocrine-Disrupting Chemicals Use Distinct Mechanisms of Action to Modulate Endocrine System Function. *Endocrinology* **2006**, *147* (6), s25−s32.

(5) Diamanti-Kandarakis, E.; Palioura, E.; Kandarakis, S. A.; Koutsilieris, M. The Impact of Endocrine Disruptors on Endocrine Targets. *Horm. Metab. Res.* **2010**, *42* (8), 543−552.

(6) Schug, T. T.; Janesick, A.; Blumberg, B.; Heindel, J. J. Endocrine Disrupting Chemicals and Disease Susceptibility. *J. Steroid Biochem. Mol. Biol.* **2011**, *127* (3−5), 204−215.

(7) Moral, R.; Wang, R.; Russo, I. H.; Lamartiniere, C. A.; Pereira, J.; Russo, J. Effect of Prenatal Exposure to the Endocrine Disruptor Bisphenol A on Mammary Gland Morphology and Gene Expression Signature. *J. Endocrinol.* **2007**, *196* (1), 101−112.

(8) Ung, C. Y.; Li, H.; Yap, C. W.; Chen, Y. Z. In Silico Prediction of Pregnane X Receptor Activators by Machine Learning Approaches. *Mol. Pharmacol.* **2006**, *71*, 158.

(9) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37* (1−2), 1700153.

(10) Merk, D.; Grisoni, F.; Friedrich, L.; Gelzinyte, E.; Schneider, G. Computer-Assisted Discovery of Retinoid X Receptor Modulating Natural Products and Isofunctional Mimetics. *J. Med. Chem.* **2018**, *61*, 5442.

(11) Merk, D.; Grisoni, F.; Friedrich, L.; Gelzinyte, E.; Schneider, G. Scaffold Hopping from Synthetic RXR Modulators by Virtual Screening and de Novo Design. *MedChemComm* **2018**, *9*, 1289.

(12) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124* (7), 1023−1033.

(13) Chen, Y.; Cheng, F.; Sun, L.; Li, W.; Liu, G.; Tang, Y. Computational Models to Predict Endocrine-Disrupting Chemical Binding with Androgen or Oestrogen Receptors. *Ecotoxicol. Environ. Saf.* **2014**, *110*, 280−287.

(14) Trisciuzzi, D.; Alberga, D.; Mansouri, K.; Judson, R.; Novellino, E.; Mangiatordi, G. F.; Nicolotti, O. Predictive Structure-Based Toxicology Approaches to Assess the Androgenic Potential of Chemicals. *J. Chem. Inf. Model.* **2017**, *57* (11), 2874−2884.

(15) OECD (Organization for Economical Cooperation and Development). Guidance Document on the Validation of (Q)SAR Models; ENV/JM/MONO(2007)2; OECD Environment Health and Safety Publications: Series on Testing and Assessment, 2007.

(16) Kavlock, R.; Chandler, K.; Houck, K.; Hunter, S.; Judson, R.; Kleinstreuer, N.; Knudsen, T.; Martin, M.; Padilla, S.; Reif, D.; et al. Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management. *Chem. Res. Toxicol.* **2012**, *25* (7), 1287−1302.

(17) Tice, R. R.; Austin, C. P.; Kavlock, R. J.; Bucher, J. R. Improving the Human Hazard Characterization of Chemicals: A Tox21 Update. *Environ. Health Perspect.* **2013**, *121* (7), 756.

(18) Kleinstreuer, N. C.; Ceger, P.; Watt, E. D.; Martin, M.; Houck, K.; Browne, P.; Thomas, R. S.; Casey, W. M.; Dix, D. J.; Allen, D.; et al. Development and Validation of a Computational Model for Androgen Receptor Activity. *Chem. Res. Toxicol.* **2017**, *30* (4), 946−964.

(19) Kode srl. *Dragon (Software for Molecular Descriptor Calculation) Version 7.0*; 2016. https://Chm.Kode-Solutions.Net (accessed Jan 31, 2019).

(20) ScrubChem. https://www.Scrubchem.Org (accessed Jan 31, 2019).

(21) Mansouri, K.; Kleinstreuer, N. C.; Watt, E. D.; Harris, J.; Judson, R. *CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity*; Poster. 2017. https://Doi.Org/10.23645/Epacomptox.5176876.V1 (accessed Jan 31, 2019).

(22) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics (2 Vols.)*; Wiley-VCH Verlag GmbH: Weinheim, Germany: Weinheim, 2009; Vol. 41.

(23) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(24) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38* (19), 2894−2896.

(25) Todeschini, R.; Ballabio, D.; Cassotti, M.; Consonni, V. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *J. Chem. Inf. Model.* **2015**, *55* (11), 2365−2374.

(26) Dasarathy, B. V. *Nearest Neighbor NN Norms: NN Pattern Classification Techniques*; 1991.

(27) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123−140.

(28) Rokach, L.; Maimon, O. Classification Trees. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer US: Boston, MA, 2010; pp 149−174,.

(29) Grisoni, F.; Consonni, V.; Vighi, M.; Villa, S.; Todeschini, R. Investigating the Mechanisms of Bioconcentration through QSAR Classification Trees. *Environ. Int.* **2016**, *88*, 198−205.

(30) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer series in statistics, New York, NY, USA, 2001; Vol. *1*, DOI: 10.1007/978-0-387-84858-7.

(31) Cestnik, B. Estimating Probabilities: A Crucial Task in Machine Learning. In *ECAI*; 1990; Vol. *90*, pp 147−149.

(32) McCallum, A.; Nigam, K.; others. A Comparison of Event Models for Naive Bayes Text Classification. In *AAAI-98 workshop on learning for text categorization*; Citeseer: 1998; Vol. *752*, pp 41−48.

(33) Chen, S. F.; Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling. *Comput. Speech Lang.* **1999**, *13* (4), 359−394.

(34) Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate Comparison of Classification Performance Measures. *Chemom. Intell. Lab. Syst.* **2018**, *174*, 33−44.

(35) MathWorks Inc. MATLAB R2017a. mathworks.com (accessed Sep 6, 2017).

(36) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME - the Konstanz Information Miner: Version 2.0 and Beyond. *SIGKDD Explor Newsl* **2009**, *11* (1), 26−31.

(37) Batagelj, V.; Mrvar, A. Pajek-Program for Large Network Analysis. *Connections* **1998**, *21* (2), 47−57.

(38) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17* (5), 4791−4810.

(39) Ballabio, D.; Biganzoli, F.; Todeschini, R.; Consonni, V. Qualitative Consensus of QSAR Ready Biodegradability Predictions. *Toxicol. Environ. Chem.* **2016**, *99* (7−8), 1193−1216.

(40) Papa, E.; van der Wal, L.; Arnot, J. A.; Gramatica, P. Metabolic Biotransformation Half-Lives in Fish: QSAR Modeling and Consensus Analysis. *Sci. Total Environ.* **2014**, *470−471*, 1040−1046.

(41) Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50−60.

(42) Pereira de Jésus-Tran, K.; Côté, P.-L.; Cantin, L.; Blanchet, J.; Labrie, F.; Breton, R. Comparison of Crystal Structures of Human Androgen Receptor Ligand-Binding Domain Complexed with Various Agonists Reveals Molecular Determinants Responsible for Binding Affinity. *Protein Sci.* **2006**, *15* (5), 987−999.

(43) Ullrich, T.; Sasmal, S.; Boorgu, V.; Pasagadi, S.; Cheera, S.; Rajagopalan, S.; Bhumireddy, A.; Shashikumar, D.; Chelur, S.; Belliappa, C.; et al. 3-Alkoxy-Pyrrolo[1,2-b]Pyrazolines as Selective Androgen Receptor Modulators with Ideal Physicochemical Properties for Transdermal Administration. *J. Med. Chem.* **2014**, *57* (17), 7396−7411.

(44) 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* (21), 7160−7174.

(45) Prim, R. C. Shortest Connection Networks and Some Generalizations. *Bell Syst. Tech. J.* **1957**, *36* (6), 1389−1401.

(46) Nembri, S.; Grisoni, F.; Consonni, V.; Todeschini, R. Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9. *Int. J. Mol. Sci.* **2016**, *17* (6), 914.

(47) Grisoni, F.; Consonni, V.; Vighi, M. Detecting the Bioaccumulation Patterns of Chemicals through Data-Driven Approaches. *Chemosphere* **2018**, *208*, 273−284.

(48) Fuchs, J. A.; Grisoni, F.; Kossenjans, M.; Hiss, J. A.; Schneider, G. Lipophilicity prediction of peptides and peptide derivatives by consensus machine learning. *MedChemComm* **2018**, *9* (9), 1538−1546.

(49) Bohl, C. E.; Miller, D. D.; Chen, J.; Bell, C. E.; Dalton, J. T. Structural Basis for Accommodation of Nonsteroidal Ligands in the Androgen Receptor. *J. Biol. Chem.* **2005**, *280* (45), 37747−37754.

(50) Tan, M. E.; Li, J.; Xu, H. E.; Melcher, K.; Yong, E. Androgen Receptor: Structure, Role in Prostate Cancer and Drug Discovery. *Acta Pharmacol. Sin.* **2015**, *36* (1), 3−23.