JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

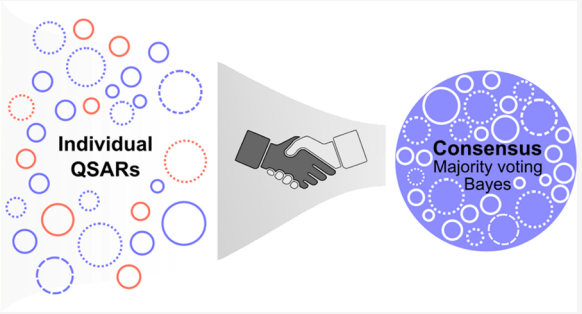
Consensus versus Individual QSARs in Classification: Comparison on a Large-Scale Case Study

Cecile Valsecchi, Francesca Grisoni, Viviana Consonni, and Davide Ballabio*

Cite This: J. Chem. Inf. Model. 2020, 60, 1215−1223 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Consensus strategies have been widely applied in many different scientific fields, based on the assumption that the fusion of several sources of information increases the outcome reliability. Despite the widespread application of consensus approaches, their advantages in quantitative structure−activity relationship (QSAR) modeling have not been thoroughly evaluated, mainly due to the lack of appropriate large-scale data sets. In this study, we evaluated the advantages and drawbacks of consensus approaches compared to single classification QSAR models. To this end, we used a data set of three properties (androgen receptor binding, agonism, and antagonism) for approximately 4000 molecules with predictions performed by more than 20 QSAR models, made available in a large-scale collaborative project. The individual QSAR models were compared with two consensus approaches, majority voting and the Bayes consensus with discrete probability distributions, in both protective and nonprotective forms. Consensus strategies proved to be more accurate and to better cover the analyzed chemical space than individual QSARs on average, thus motivating their widespread application for property prediction. Scripts and data to reproduce the results of this study are available for download.



1. INTRODUCTION

Consensus approaches aim to combine and integrate information derived from different sources to increase the outcome reliability and overcome limitations of single approaches.[1] In the framework of quantitative structure−activity relationships (QSARs), they are generally recognized as valuable tools to reduce the effects of underestimating uncertainties in the prediction of biological activities.[2,3]

The main underlying assumption of consensus modeling in QSAR is that individual models, due to their reductionist nature, consider only partial structure−activity information, as encoded by molecular descriptors and adopted algorithms. Thus, the combination of multiple QSAR predictions may provide a wider knowledge and increase the reliability associated with the predictions compared to individual models.[1,4] Indeed, one of the advantages of the consensus methods is the reduction of the effects of contradictory information by averaging the predictions of models,[1,5−8] although this is not always reflected in improvements of the predictive ability compared to single models.[1,5] Furthermore, integrating individual QSARs can broaden the applicability domain, that is, the chemical space where predictions can be considered reliable.[9,10] For these reasons, consensus methods, also known as high-level data fusion or ensemble approaches, have been extensively applied in QSAR studies.[11−16] Recent studies on the improvement achieved with large-scale consensus approaches for quantitative (regression) models can be found in the literature.[17,18] However, to the best of our knowledge, no thorough evaluation of the consensus versus single qualitative (classification) model performance has been carried out to date, since only a few QSAR models are usually available for the same endpoint.[6,10,19−23]

The present study was based on the outcome of a large collaborative project (Collaborative Modeling Project of Androgen Receptor Activity, CoMPARA[19]), which produced three data sets containing experimental values on androgen receptor (AR) modulation and corresponding QSAR predictions, namely, (i) binding to AR (34 QSAR models), (ii) AR antagonism (22 QSAR models), and (iii) AR agonism (21 QSAR models).[19] CoMPARA was chosen as a test system due to the large availability of diverse QSAR-based predictions. Note that in the framework of CoMPARA, two ad hoc consensus approaches were applied by combining predictions with a weighting score based on the goodness-of-fit, predictivity, and robustness of models.[24] However, the aim of the present study is not a comparison with these former consensus approaches, which were specifically targeted to screen and prioritize chemicals for endocrine activity, but the

Received: November 13, 2019
Published: February 19, 2020

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

systematic investigation of the advantages of further consensus strategies compared to single QSAR models. To this end, approaches with varying levels of complexity (majority voting and Bayesian methods, in both protective and nonprotective versions) were considered. Moreover, we investigated whether the exclusion of the worst-performing models may influence the consensus outcome, in terms of chemical space coverage and predictive performance.[13,15] Finally, a structural similarity analysis was carried out to identify specific chemical regions where individual QSAR models, and the respective consensus outcome, fail in their predictions.

2. MATERIALS AND METHODS

2.1. Collaborative Project. The QSAR models considered in this work were previously developed in the framework of a collaborative project (Collaborative Modeling Project of Androgen Receptor Activity, CoMPARA[24]), coordinated by the National Center of Computational Toxicology (U.S. Environmental Protection Agency). CoMPARA aimed to develop in silico approaches to identify potential androgen receptor (AR) modulators. This project involved 25 research groups worldwide, which were provided with a calibration set consisting of 1689 chemicals and the corresponding experimental annotations on binding, agonism, and antagonism activities (in the form of qualitative labels, active/inactive), as determined by a battery of 11 in vitro assays.[20] The research groups were then asked to predict another 55 450 chemicals for one or more endpoints (binding, agonism, and antagonism) using their own developed QSAR models. Finally, these predictions were merged through ad hoc consensus approaches, which are currently being used by the CoMPARA coordinators to prioritize experimental tests for potential endocrine-disrupting chemicals.[24]

The predictive ability of individual QSAR models was assessed by the project coordinators on the basis of three specific evaluation sets, which were embedded within the large prediction set of 55 450 chemicals, to carry out a blinded verification. These sets were created from literature data extracted from different sources and curated for quality, by considering target, modality, hit call, and concordance among the annotated values. The three evaluation sets included 3540 chemicals annotated with binding activities, 4408 with agonism, and 3667 with antagonism. We used the individual QSAR predictions for these three evaluation sets, whose details are summarized in Table 1, to calculate the consensus approaches. All evaluation sets are characterized by unbalanced sample distribution toward inactivity with 88.4, 91.4, and 96.3% of inactive chemicals for binding, antagonism, and agonism, respectively. The three evaluation sets, including chemical identifiers, SMILES, and predictions, are available as the Supporting Information describing the CoMPARA project.[24]

Note that although the project coordinators also provided quantitative binding, agonism, and antagonism activities, the participants developed only a few regression models (five, five, and three for binding, agonist, and antagonist, respectively). We considered, thus, only classification models for consensus approaches to allow for a comprehensive and systematic analysis.

2.2. Individual QSAR Models. CoMPARA consortium members trained QSAR models to classify chemicals for their potential of AR binding (34 models), agonism (21 models), and antagonism (22 models). Models were mainly developed on the same calibration set of 1689 chemicals, using different modeling strategies (e.g., artificial neural networks, *k*-nearest neighbors, support-vector machines, partial least squares discriminant analysis, classification trees[8,22,23]) and molecular descriptors (e.g., binary fingerprints and nonbinary descriptors).[24] Each submitted prediction was associated with the applicability domain (AD) assessment, that is, an indication on whether predictions can be considered as reliable.[9,25]

The predictive ability of QSAR models was assessed on the evaluation set through the following classification measures: (i) sensitivity (Sn) and specificity (Sp), which are the percentages of correctly classified active and inactive chemicals, respectively, and (ii) the non-error rate (NER), also known as balanced accuracy, that is the average of Sn and Sp.[26] Moreover, the percentage of reliably predicted chemicals (coverage, Cvg) was used as an additional criterion to assess the model performances. The distribution of the classification estimators of the individual CoMPARA models for the three modeled endpoints is summarized in Figure 1.

All models have a good predictive performance, with the median NER ranging from 71.0% (antagonism) to 83.8% (agonism). Specificity values (Sp) are always higher than sensitivities (Sn), thus indicating a better performance of the models in the identification of inactive compounds. Except for the agonism endpoint, sensitivity is associated with a higher variability than specificity, with values ranging from ∼20 to ∼80% on both binding (relative standard deviation equal to ∼28%) and antagonism (relative standard deviation equal to ∼29%) endpoints. This general behavior can be due to both unbalanced classes, which are strongly skewed toward inactivity (88.4 and 91.4% of inactive molecules for binding and antagonism data sets, respectively; Table 1), and differences in the ranges of testing between training and evaluation sources, as reported in the literature.[24]

The models for agonism show the best trade-off between sensitivity (Sn) and specificity (Sp), with most models characterized by sensitivity values in the range of ∼70 to ∼84% and specificity in the range of ∼76 to ∼100%. Additionally, agonism models have the highest median sensitivity (76.2%), specificity (96.3%), and NER (83.8%), although the agonism data set includes only 3.7% of actives and is thus the most unbalanced among the three evaluation sets (see Table 1). Models for binding and antagonism have similar median NERs (74.8 and 71%, respectively), moderately low median sensitivities (64.1 and 55.9%), and high median specificities (88.3 and 85.5%).

The majority of individual models are characterized by a high percentage of reliably predicted chemicals (coverage values equal to 88.1, 88.1, and 89.5% on average for binding, antagonism, and agonism, respectively). The models that are

Table 1. Number of Chemicals (Total, Actives, and Inactives) Included in the CoMPARA Binding, Antagonism, and Agonism Evaluation Sets and Number of Models Developed within the CoMPARA Project for Each Endpoint

	binding	antagonism	agonism
number of chemicals	3540	3667	4408
active	411 (11.6%)	314 (8.6%)	164 (3.7%)
inactive	3129 (88.4%)	3353 (91.4%)	4244 (96.3%)
individual QSAR models	34	22	21

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

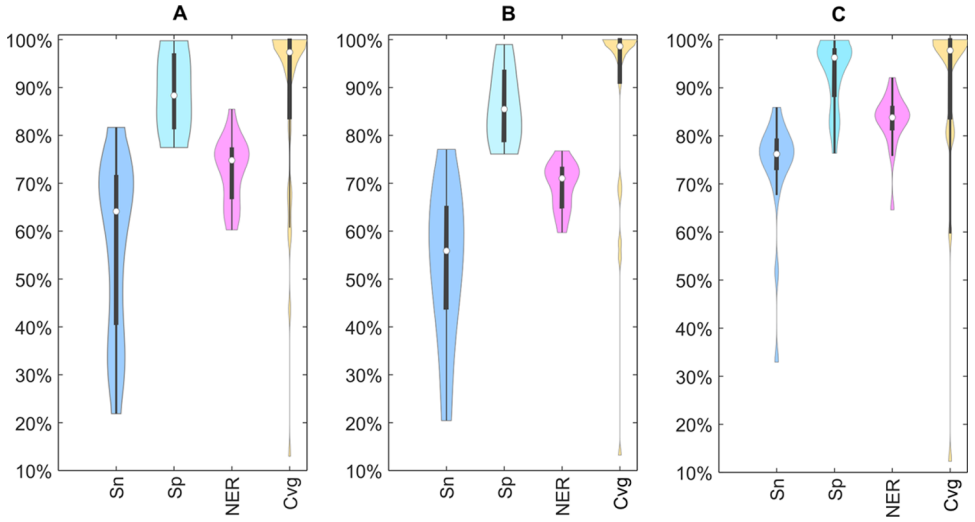


Figure 1. Violin plots of sensitivity (Sn), specificity (Sp), non-error rate (NER), and coverage (Cvg) for the individual CoMPARA models on the binding (A), antagonism (B), and agonism (C) evaluation sets. Empty dots indicate median values, thick gray lines indicate the second and third quartiles, and thin gray lines indicate the first and fourth quartiles. Shapes indicate the underlying data distribution. Numerical values of the classification parameters for all of the models are provided in the Supporting Information (Tables S1−S3).

able to reliably predict only a few molecules are associated with the highest classification performance, thus confirming that high classification performance is more likely on a narrow applicability domain. In fact, the four best models to predict the binding activity (NER higher than 80%) were characterized by a limited percentage of chemicals in their applicability domain (coverage values equal to 13, 43.7, 60.7, and 69%; Table S1), suggesting that these single models have limited applications for prioritization purposes.

2.3. Consensus Methods. In this study, two consensus strategies were applied to integrate the predictions provided by individual models: majority voting and the Bayes consensus with discrete probability distributions. These methods are briefly described below.

2.3.1. Majority Voting. Voting methods combine the predictions provided by independent models with different frequency-based strategies, such as averaging and scoring.[14,16,23,27] The most simple and intuitive voting approach is the majority voting (MV) rule, which assigns a chemical to the most frequently predicted class among the pool of considered models.[28,29] Cautionary (protective) voting approaches can be obtained by considering only predictions integrated with a sufficiently high concordance (based on a user-defined threshold) among the pool of models.

In this work, we considered three different majority voting strategies as follows: (i) majority voting loose (MVL), (ii) majority voting intermediate (MVI), and (iii) majority voting strict (MVS). The "loose" approach classifies molecules using the most recurrent class assignment. In the two-class case, this corresponds to the class predicted with a frequency higher than 50%. The "intermediate" and "strict" criteria (MVI and MVS, respectively) are protective approaches. MVS assigns the compound only if the prediction agreement is higher than or equal to 75%. The MVS approach provides a prediction for a given molecule only if all of the individual models predict the same class (100% agreement). To ensure the reliability of the consensus outcome, only the predictions within the applicability domain of individual models were considered for the calculation of the agreement.

2.3.2. Bayesian Consensus. An alternative to the majority voting approach is a probabilistic method, such as Bayesian consensus. The Bayes rule,[12,30,31] in particular, estimates the prior probability for a molecule to belong to a specific class for each information source and then combines this information to provide a joint probability.[32]

In particular, the Bayes consensus with discrete probability distributions[31,33] initially takes into account the first evidence, e, which is in this case the class (active or inactive) predicted by the first model. Then, the posterior probabilities $p(h_g|e)$ that hypothesis h_g is true given evidence e are calculated for any class g, as follows

$$p(h_g|e) = \frac{p(e|h_g)\cdot p(h_g)}{\sum_g p(e|h_g)\cdot p(h_g)} \quad (1)$$

where $p(e|h_g)$ is the likelihood probability that evidence e is observed given that hypothesis h_g is true and $p(h_g)$ is the prior probability that hypothesis h_g is true in the absence of any specific evidence.

With two hypotheses (i.e., class equal to "active" or "inactive"), the prior equal (noninformative) probability is estimated as $p(h_{\mathrm{ACTIVE}}) = p(h_{\mathrm{INACTIVE}}) = 0.50$. The prior proportional (informative) probability for each hypothesis h_g would be $p(h_g) = n_g/n$, where n_g is the number of molecules belonging to the gth experimental class within the n total molecules.

Likelihood probabilities for each model can be estimated from its confusion matrix, where the numbers of correct and incorrect classifications are collected.[31] Once posterior probabilities for the first model have been calculated, the Bayes consensus proceeds with the following iterative procedure. Posterior probabilities of the first model are used as new prior probabilities for the second step, where the class predicted by the second model is the new evidence e on the basis of which the posterior probabilities are calculated. These posterior probabilities become the new prior probabilities in the third iteration and so on, until predictions of all models have been used in the consensus process. At the end of the iterations, the posterior probabilities corresponding to the

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

Table 2. Classification Performance of the Consensus Approaches for Binding, Agonism, and Antagonism Endpoints[aa]

	binding (34 models)					antagonism (22 models)					agonism (21 models)				
consensus approach	Sn (%)	Sp (%)	NER (%)	Cvg (%)	rank	Sn (%)	Sp (%)	NER (%)	Cvg (%)	rank	Sn (%)	Sp (%)	NER (%)	Cvg (%)	rank
MVL	61.8	91.8	76.8	99.3	4	61.5	87.3	74.4	98.9	3	73.8	97.5	85.7	99.7	2
MVI	60.6	98.3	79.5	80.6	8	60.0	93.8	76.9	80.1	4	76.1	99.0	87.5	91.5	6
MVS	26.9	100	63.5	37.5	39	39.0	99.2	69.1	42.4	25	64.8	99.9	82.3	51.4	17
B	72.3	84.9	78.6	100	1	71.0	81.2	76.1	100	1	74.4	95.1	84.7	100	3
Bp	73.3	85.9	79.6	96.1	7	73.5	82.9	78.2	92.9	2	75.8	95.9	85.9	97.7	4

[a]For each consensus approach, sensitivity (Sn), specificity (Sp), non-error rate (NER), coverage (Cvg), and total ranking are reported. B, Bayes; Bp, protective Bayes; MVL, majority voting loose; MVI, majority voting intermediate; MVS, majority voting strict.

combination of all of the information sources are obtained. Therefore, the Bayes consensus assigns a probability value to each class, which is then used for prediction, by choosing the class with the maximum posterior probability. As for the majority voting strategies, the Bayes consensus can be used in a protective manner by setting a posterior probability threshold (in this study, 95%) that has to be fulfilled to predict the class.[31]

When proportional prior probabilities are used with models calibrated on data with unbalanced class distributions, Bayes results may change depending on which model sequence enters the iteration process. In fact, if models are associated with different prior proportional probabilities, the model entering the first position of the iterative process can produce a different outcome with respect to others. This is the case of the collaborative project under analysis, whose models were calibrated and validated on the same set of chemicals, but with different ratios of molecules included in the applicability domain, leading to different prior probabilities. To overcome this potential issue, in this study, we used equal prior probabilities.[11]

As for the majority voting approaches, predictions associated with molecules outside the applicability domain of individual models were not considered.

2.4. Analysis of Molecular Similarities. A molecular similarity analysis was carried out to investigate potential relationships between the molecular structure and misclassifications provided by QSAR and consensus models. To this end, extended connectivity fingerprints (ECFPs),[34] which encode for the presence of branched substructures in a binary array, were used as molecular descriptors, with the setting specified in Section 2.5. Pairwise molecular similarities, as quantified using the Jaccard−Tanimoto similarity coefficient,[35] were used to produce a two-dimensional representation of the molecular space by means of multidimensional scaling (MDS).[36]

2.5. Software. ECFP04 (1024 bits and 0−2 bond radius) were calculated by means of DRAGON 7[37] with default settings ("Bits per pattern" = 2; "Count fragments": True; "Atom Options": [Atom type, Aromaticity, Connectivity total, Charge, Bond order]). MDS was carried out in MATLAB 2018b[38] by a publicly available toolbox.[39] Consensus strategies were performed using the MATLAB code written by the authors, which is available for download at http://www.michem.unimib.it/download/data/bayes-and-majority-voting-consensus-for-matlab/. Violin plots were created with the code available at the URL https://github.com/bastibe/Violinplot-Matlab.

3. RESULTS

3.1. Analysis of Consensus Strategies. *3.1.1. Classification Performance.* The selected consensus strategies (i.e., Bayes [B], protective Bayes [Bp], majority voting loose [MVL], majority voting intermediate [MVI], and majority voting strict [MVS]) were used to integrate the predictions of the individual QSAR models for binding, antagonism, and agonism. When applying protective consensus strategies, the outcome predictions were rejected if related to potential uncertainty, that is, (i) prediction agreement lower than 75 and 100% for MVI and MVL, respectively, and (ii) posterior probability lower than 95% for protective Bayes. For majority voting loose (MVL), no prediction was provided in the case of equal frequency for the two classes (50%).

In analogy with the individual models, the consensus approaches were evaluated for their classification performance, in terms of sensitivity (Sn), specificity (Sp), non-error rate (NER), and coverage (Cvg) (Table 2). A graphical comparison with individual models is represented in Figure 2 with plots of sensitivity versus specificity values. Moreover, since sensitivity, specificity, and coverage have the same unit scale and optimality direction (i.e., ranging from 0 to 100%; the closer to 100%, the better), a comprehensive performance index was calculated as their arithmetic average, denoted as "Utility" in the framework of ranking analysis and multicriteria decision making.[40−43] Both consensus and individual QSARs were ranked for decreasing values of Utility (Table 2).

Consensus strategies have better NERs than individual QSARs on average, without substantial losses in terms of coverage compared to individual models; additionally, consensus models are always ranked among the top 10 positions (Table 2). The exception is MVS, which provides a remarkably lower coverage (lower than 52% for all of the endpoints), due to the required 100% agreement among multiple predictions (up to 34 predictions). The narrow coverage of MVS, however, was not counterbalanced by a better performance compared to the other consensus approaches. MVS, in fact, showed the lowest NER and Cvg values among all of the tested consensus strategies. For these reasons, MVS was not analyzed further in this framework.

Unlike MVS, the other consensus strategies generally showed a better trade-off between the classification performance and the chemical space coverage than individual QSARs. For instance, the two single-binding models in the upper-right region of the sensitivity versus specificity space (Figure 2A) have the best predictive performance for binding, with NERs equal to 85.5 and 83.8%, respectively (Table S1), but they cover only a small portion of the chemical space, as it results from the small coverage values (43.7 and 60.7%, respectively).

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

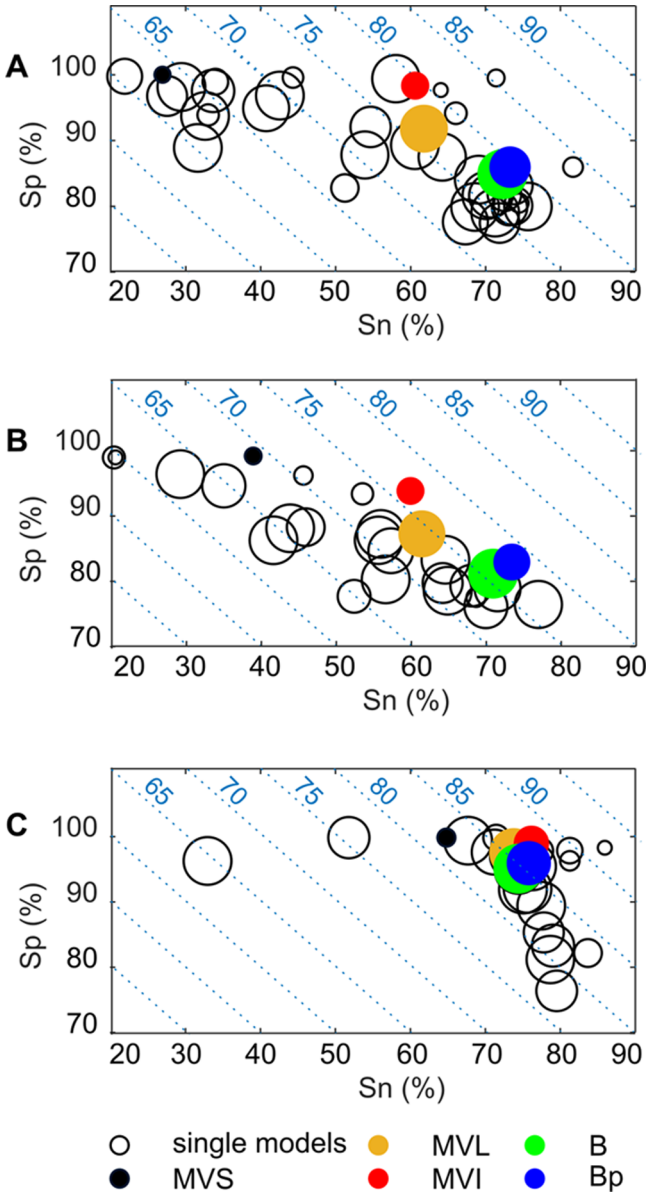


Figure 2. Plot of sensitivities (Sn) versus specificities (Sp) for the individual models (black empty circles) and for the consensus approaches (filled, colored circles) for each endpoint: (A) binding, (B) antagonism, and (C) agonism. Green, blue, red, yellow, and black circles indicate Bayes (B), Bayes protective (Bp), majority voting intermediate (MVI), loose (MVL), and strict (MVS) consensus, respectively. The size of the circles is proportional to the coverage (Cvg); the smaller a circle, the lower the coverage. Isolines represent NER variations (5% steps).

On the other hand, the protective Bayes (Bp) reached a slightly lower NER (79.6%) but higher coverage (96.1%).

The models on binding and antagonism (Figure 2A,B) endpoints are characterized by the unbalanced specificity and sensitivity values, with several models showing high specificity (Sp > 90%) and low sensitivity (Sn < 50%). For these endpoints, consensus methods achieved more balanced values of sensitivity and specificity, due to the compensation in the integration of diverse sources of information. This is particularly evident in the case of the Bayes approaches (Table 2), ranked as the best overall approach for both binding and antagonism, and confirms that the uncertainty can be reduced by the integration of conflicting sources.

The difference in the performance between consensus and individual QSARs is less pronounced when considering agonism (Figure 2C), since the individual models have more homogeneous NERs and balanced Sn and Sp values compared to the other case studies. Therefore, consensus methods converged to similar performances.

Majority voting approaches inherit the high specificity values of individual models for both binding and antagonism endpoints, while the Bayes consensus led to a higher sensitivity. This trend could be caused by the low false-positive rates of individual models (Figure 1) and the way this information is weighted and integrated into the Bayes calculation (eq 1). Thus, in this framework, if a compound is predicted with an equal frequency as active and inactive by the individual models, it will be more likely assigned to the active class by the Bayes consensus.

Protective approaches (MVI and Bp) yielded slightly better results in terms of the classification performance (NER) compared to their nonprotective counterparts, but with a relatively larger loss in coverage (up to 18.7% loss), especially when dealing with majority voting schemes. This explains the worse position within the ranking of protective approaches with respect to nonprotective ones (Table 2). As an example, the MVL approach on the binding endpoint led to an NER of 76.8% and a coverage of 99.3% (rank 4), while the protective MVI led to a slightly higher NER (79.5%) but considerably lower coverage (80.6%) and a worse rank (8).

3.1.2. Chemical Space Analysis. To evaluate potential associations between misclassifications and structural chemical features, compounds were described by extended connectivity fingerprints (ECFPs). A multidimensional scaling (MDS) was then performed to visualize the similarity relationships (as encoded by the Jaccard−Tanimoto similarity coefficients calculated on ECFPs) in a bidimensional plot. This allowed us to analyze the relationship between such a structural representation and the number of models (individual or consensus), providing reliable predictions.

In the obtained MDS representation (Figure 3 for the binding endpoint), chemicals are arranged in two clusters. The cluster characterized by negative scores on the first dimension is mainly composed of aliphatic molecules with long alkyl chains, as well as cyclic aliphatic compounds, mostly with sp^3-hybridized carbon atoms. The most frequent functional groups are carbonyls, hydroxyls, ethers, and esters, while conjugated structures or p-systems are almost absent in this cluster. The second cluster, located in the positive score region on the first dimension, is mainly composed of conjugated structures, primarily aromatic rings with many electron acceptor substituents (e.g., $-NO_2$, $-PO_3$, $-SO_3$, −F, −Cl, and −CO) and a few donating groups (e.g., $-NH_2$ and −OH).

Most of the misclassified molecules cluster in specific regions of the chemical space. Similar distributions were obtained for agonism and antagonism data sets (see Figures S1 and S2). Aliphatic chemicals (characterized by negative scores on the first dimension) are in general well-predicted; on the other hand, misclassifications seem to be mainly grouped in the aromatic cluster (positive scores on the first dimension). Besides incorrect predictions, this region is also associated with lower coverage of the individual models (Figure 3A). Similarly, the intermediate region between the two clusters is characterized by low coverage, reflecting regions of model

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

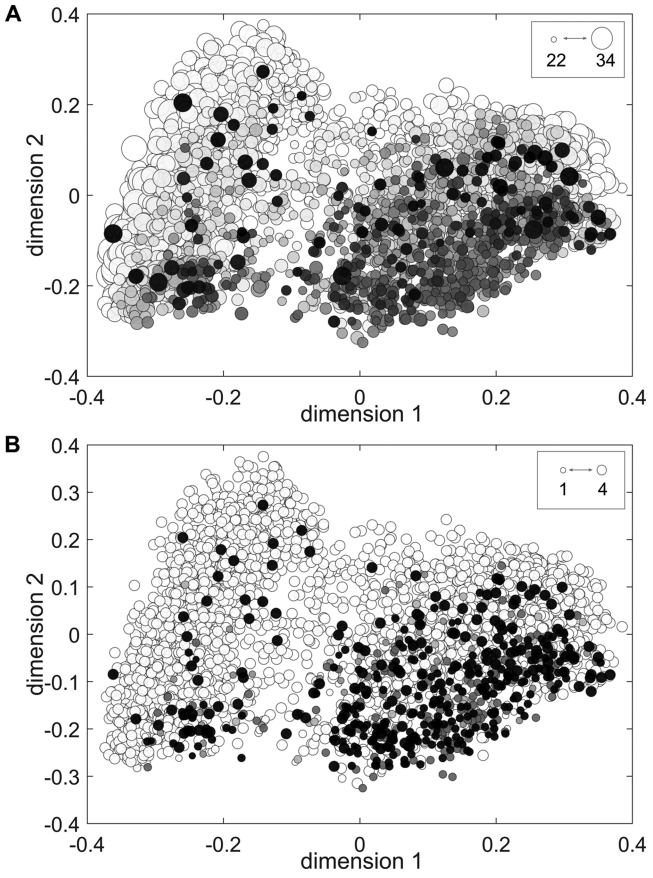


Figure 3. Plot of the first and second dimensions of the MDS for the binding endpoint (ECFPs, Jaccard−Tanimoto similarity). Each point represents a chemical, colored based on the number of misclassifications of (A) individual QSAR models and (B) consensus strategies; the darker the point, the higher the number of misclassifications. The size of each point is proportional to the percentage of models or consensus strategies (A and B, respectively) that provided a prediction for the chemical.

uncertainty. These observations point toward the presence of relationships between chemical features (as encoded within ECFPs) and model performances, since misclassifications are mainly located in limited portions of the chemical space, where molecules are often out of the models' applicability domains.

Some chemicals were incorrectly classified by all of the individual QSAR models despite being in their applicability domain, as follows: 19 molecules for binding (all false negatives), 28 for agonism (25 false negatives and 3 false positives), and 37 for antagonism (25 false negatives and 12 false positives). We identified some recurring issues that might explain the observed misclassifications:

1. *Borderline Compounds.* Several active molecules that were consistently predicted as inactive are labeled as having experimental weak or very weak potency (Table S4), as quantified by the half-maximal activity (AC_{50}, the molar concentration that produces 50% of the maximum possible activity). The molecules were thus labeled as active, but they actually are borderline between activity and inactivity. Additionally, different activity values due to differences among experimental protocols have been already reported on this set of chemicals.[24] In such cases, models and experimental data can be regarded as belonging to the same level of assessment[44] and QSAR models might provide an indication of the potential inactivity of these consistently misclassified compounds.
2. *Differences between Charged and Neutralized Forms.* Another reason could be related to the different activities of charged compounds toward their neutralized counterparts. In fact, traditional QSAR pipelines do not consider annotated counterions and rely on the neutralized form for descriptor calculations. Nine false negatives (two, one, and six for binding, antagonism, and agonism sets, respectively) showed a different activity in their neutralized form and with an annotated counterion (Table S4). For example, 1-butyl-4-methylpyridinium hexafluorophosphate (DTXSID4049296, CASRN 401788-99-6) is a moderate antagonist (AC_{50} = 1.94

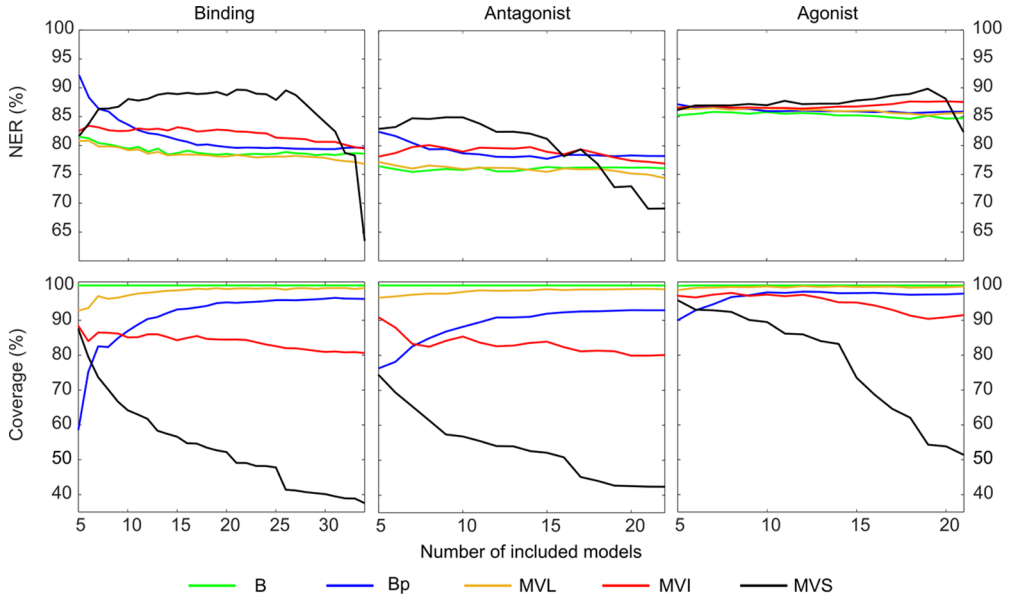


Figure 4. Plot of NER and coverage as a function of the number of models included in the consensus calculation. B, Bayes; Bp, protective Bayes; MVL, majority voting loose; MVI, majority voting intermediate; MVS, majority voting strict.

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

Table 3. Classification Performance of the Consensus Approaches Estimated on the Binding, Antagonism, and Agonism Sets Considering the Best Five Models Only (Selected Based on NER)[a]

	binding (5 models)					antagonism (5 models)					agonism (5 models)				
consensus approach	Sn (%)	Sp (%)	NER (%)	Cvg (%)	rank	Sn (%)	Sp (%)	NER (%)	Cvg (%)	rank	Sn (%)	Sp (%)	NER (%)	Cvg (%)	rank
MVL	63.9	97.7	80.8	92.8	3	71.6	82.8	77.2	96.5	2	73.8	98.8	86.3	98.6	2
MVI	65.7	99.3	82.5	88.4	4	71.9	84.4	78.1	90.9	3	74.1	99.0	86.5	97.0	4
MVS	63.8	99.5	81.6	87.4	6	78.3	87.6	83.0	74.4	5	73.1	99.2	86.1	95.8	6
B	72.0	91.0	81.5	100	1	73.2	79.7	76.5	100	1	74.4	96.1	85.2	99.9	3
Bp	88.3	96.2	92.2	58.6	7	79.4	85.5	82.4	76.3	4	76.1	98.2	87.1	90.0	7

[a]For each consensus approach, sensitivity (Sn), specificity (Sp), non-error rate (NER), coverage (Cvg), and total ranking are reported. B, Bayes; Bp, protective Bayes; MVL, majority voting loose; MVI, majority voting intermediate; MVS, majority voting strict.

μM), but its neutralized form (with removed counterion) is identical to the neutralized forms of 1-butyl-4-methylpyridinium bromide (DTXSID2049345, CASRN 65350-59-6) and 1-butyl-4-methylpyridinium trifluoromethanesulfonate (DTXSID5049368, CASRN 882172-79-4), which are inactive. This highlights the need for considering the effect of charge and counterions on the final biological activity.

Although consensus methods reduced the uncertainty (Figure 3B), misclassifications and unclassified chemicals are still mainly located in the critical region characterized by positive scores of the MDS space (aromatic cluster), thus following the same pattern as individual models. This confirms that consensus approaches can reduce uncertainty but cannot remove it since the integration of erroneous information leads anyway to poor predictions. The performance of consensus models could improve by considering the structural features of chemicals and the individual models' performance in the chemical space.

3.2. Consensus Based on Subsets of Models. When integrating several sources of information, one could decide to select only the most reliable ones aiming to neglect misleading information and potentially improve the prediction performance. To this end, we investigated the performance of consensus strategies as a function of the number of merged individual QSARs, ordered by decreasing predictive performance. For each endpoint, subsets of models were selected as inputs for the consensus approaches with the following strategy: (i) the individual QSAR models were ranked according to their NER; (ii) consensus approaches were then calculated iteratively adding one model at a time, starting from an initial subset including the best top five (Figure 4).

The NERs of B, MVI, and MVL are slightly influenced by the number of included models. This indicates that these methods are not sensitive to the integration of poor sources of information in the consensus process. On the contrary, the protective Bayes approach (Bp) is characterized by better performances when a few good models are included, at the expense of the coverage, which shows a considerable decrease. Therefore, when the maximization of the prediction reliability is the only priority, only the most reliable sources of information shall be used in the consensus. When the final goal is to screen a large set of chemicals for testing prioritization, as in the case of the CoMPARA project, the inclusion of all of the available sources of information can considerably enhance the coverage without a significant loss of performance. MVS is the consensus approach showing the highest dependence on the number of included models; in particular, as soon as spurious information sources enter in the consensus process, the coverage significantly decreases.

Table 3 collects the classification performance of consensus approaches calculated on the top five models (chosen based on NER), which is on average better than that of individual models, with consensus strategies occupying the first seven ranking positions for all of the three considered case studies.

The protective consensus (Bp, MVI, and MVS) obtained on this reduced pool of models provided higher sensitivities than those based on the integration of all available models (Table 2), especially for binding and antagonism. However, protective approaches are always ranked worse than the nonprotective counterparts. Finally, the performance of MVS improves, since it is easier to reach a 100% prediction agreement with a few input models compared to using the whole set. For example, for binding endpoints, the NER increased from 63.5 to 81.6% and the coverage increased from 37.5 to 87.4%, respectively.

4. CONCLUSIONS

In this study, we evaluated the extent to which consensus modeling can outperform individual QSARs, by leveraging a large set of QSAR model predictions on androgen receptor binding, agonism, and antagonism. The protective and nonprotective majority voting and Bayes consensus methods were evaluated for their capability to reduce the prediction uncertainty, increase the classification performance, and overcome limitations of individual QSAR models.

The applied consensus strategies provided a better trade-off between the classification performance and the number of reliably predicted chemicals compared to single QSARs. In fact, consensus methods could correctly weigh in and integrate diverse sources of information, leading to balanced values of sensitivity and specificity, as well as to increased coverage compared to the average of individual QSARs. In fact, only a few models could perform better than consensus in terms of classification indices, but they included a limited percentage of chemicals in their applicability domain.

Protective consensus approaches were found to be suitable to incorporate information of less reliable predictions into the final assessment, thereby providing a slightly better classification performance, at the expense of the coverage.

However, consensus strategies were not able to perform well in those critical regions of the chemical space where most of the individual models failed, since the integration of erroneous information leads, by definition, to poor predictions. Implementation of a structure-driven model selection could help overcome these limitations of consensus approaches.

The performance of consensus strategies was finally evaluated as a function of the number of models included in

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223

the integration approach. The difference in terms of the classification performance between nonprotective consensus strategies applied to all of the available models and to the subset of the five most reliable ones is on average around 1% of the non-error rate (balanced accuracy). Therefore, the performance of nonprotective strategies was not significantly influenced by the presence of poorly predictive individual models, thus again demonstrating the ability of these methods to weigh in and integrate conflicting information. On the contrary, protective approaches benefit from the selection of the most predictive models.

Our final recommendation is to choose the consensus approaches based on the envisaged model application. For prioritization purposes, where one might want to predict the largest number of compounds possible, we recommend using nonprotective approaches. In this case, since MV and the Bayes consensus lead to comparable performances, MV could be the method of choice, due to the easier implementation and interpretation of the results. When the objective is, instead, to obtain the most accurate estimate possible, at the expense of the covered chemical space, protective methods should be applied on a subset of selected, best-performing models.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b01057.

> Classification performances of the individual models on the binding, agonist and antagonist evaluation sets. Summary of the molecules which were considered outside the applicability domain or misclassified by all the individual QSAR models. Plot of first and second multidimensional scaling dimensions for the agonism and antagonism sets. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Davide Ballabio − *Milano Chemometrics and QSAR Research Group, University of Milano Bicocca, 20126 Milano, Italy;* orcid.org/0000-0002-5748-147X; Email: davide.ballabio@unimib.it

Authors

Cecile Valsecchi − *Milano Chemometrics and QSAR Research Group, University of Milano Bicocca, 20126 Milano, Italy*

Francesca Grisoni − *Department of Chemistry and Applied Biosciences, ETH Zurich, 8049 Zurich, Switzerland;* orcid.org/0000-0001-8552-6615

Viviana Consonni − *Milano Chemometrics and QSAR Research Group, University of Milano Bicocca, 20126 Milano, Italy;* orcid.org/0000-0001-6252-9805

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b01057

Author Contributions

The manuscript was written with the contribution of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Kamel Mansouri for his valuable comments and feedback on the manuscript. F.G. was supported by the Swiss National Science Foundation (SNSF, Grant No. 205321_182176).

■ REFERENCES

(1) Hewitt, M.; Cronin, M. T. D.; Madden, J. C.; Rowe, P. H.; Johnson, C.; Obi, A.; Enoch, S. J. Consensus QSAR Models: Do the Benefits Outweigh the Complexity? *J. Chem. Inf. Model.* **2007**, *47*, 1460−1468.

(2) Neumann, M. B.; Gujer, W. Underestimation of Uncertainty in Statistical Regression of Environmental Models: Influence of Model Structure Uncertainty. *Environ. Sci. Technol.* **2008**, *42*, 4037−4043.

(3) Weber, C. L.; VanBriesen, J. M.; Small, M. S. A Stochastic Regression Approach to Analyzing Thermodynamic Uncertainty in Chemical Speciation Modeling. *Environ. Sci. Technol.* **2006**, *40*, 3872−3878.

(4) Jaworska, J.; Hoffmann, S. Integrated Testing Strategy (ITS) − Opportunities to Better Use Existing Data and Guide Future Testing in Toxicology. *ALTEX* **2010**, *27*, 231−242.

(5) Grisoni, F.; Consonni, V.; Villa, S.; Vighi, M.; Todeschini, R. QSAR Models for Bioconcentration: Is the Increase in the Complexity Justified by More Accurate Predictions? *Chemosphere* **2015**, *127*, 171−179.

(6) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.

(7) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.

(8) Grisoni, F.; Consonni, V.; Ballabio, D. Machine Learning Consensus To Predict the Binding to the Androgen Receptor within the CoMPARA Project. *J. Chem. Inf. Model.* **2019**, *59*, 1839−1848.

(9) 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*, 4791−4810.

(10) Hanser, T.; Barber, C.; Marchaland, J. F.; Werner, S. Applicability Domain: Towards a More Formal Definition. *SAR QSAR Environ. Res.* **2016**, *27*, 865−881.

(11) Ballabio, D.; Biganzoli, F.; Todeschini, R.; Consonni, V. Qualitative Consensus of QSAR Ready Biodegradability Predictions. *Toxicol. Environ. Chem.* **2017**, *99*, 1193−1216.

(12) Pradeep, P.; Povinelli, R. J.; White, S.; Merrill, S. J. An Ensemble Model of QSAR Tools for Regulatory Risk Assessment. *J. Cheminform.* **2016**, *8*, No. 48.

(13) Chauhan, S.; Kumar, A. SAR and QSAR in Environmental Research Consensus QSAR Modelling of SIRT1 Activators Using Simplex Representation of Molecular Structure Consensus QSAR Modelling of SIRT1 Activators Using Simplex Representation of Molecular Structure. *SAR QSAR Environ. Res.* **2018**, *29*, 277−294.

(14) Ruiz, P.; Sack, A.; Wampole, M.; Bobst, S.; Vracko, M. Integration of in Silico Methods and Computational Systems Biology to Explore Endocrine-Disrupting Chemical Binding with Nuclear Hormone Receptors. *Chemosphere* **2017**, *178*, 99−109.

(15) Asturiol, D.; Casati, S.; Worth, A. Consensus of Classification Trees for Skin Sensitisation Hazard Prediction. *Toxicol. In Vitro* **2016**, *36*, 197−209.

(16) Abdelaziz, A.; Spahn-Langguth, H.; Schramm, K.-W.; Tetko, I. V. Consensus Modeling for HTS Assays Using In Silico Descriptors Calculates the Best Balanced Accuracy in Tox21 Challenge. *Front. Environ. Sci.* **2016**, *4*, No. 2.

(17) Zakharov, A. V.; Zhao, T.; Nguyen, D.-T.; Peryea, T.; Sheils, T.; Yasgar, A.; Huang, R.; Southall, N.; Simeonov, A. Novel

Consensus Architecture To Improve Performance of Large-Scale Multitask Deep Learning QSAR Models. *J. Chem. Inf. Model.* **2019**, *59*, 4613−4624.
(18) Ambure, P.; Gajewicz-Skretna, A.; Cordeiro, M. N. D. S.; Roy, K. New Workflow for QSAR Model Development from Small Data Sets: Small Dataset Curator and Small Dataset Modeler. Integration of Data Curation, Exhaustive Double Cross-Validation, and a Set of Optimal Model Selection Techniques. *J. Chem. Inf. Model.* **2019**, *59*, 4070−4076.
(19) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124*, 1023−1033.
(20) Judson, R. S.; Magpantay, F. M.; Chickarmane, V.; Haskell, C.; Tania, N.; Taylor, J.; Xia, M.; Huang, R.; Rotroff, D. M.; Filer, D. L.; Houck, K. A.; Martin, M. T.; Sipes, N.; Richard, A. M.; Mansouri, K.; Woodrow Setzer, R.; Knudsen, T. B.; Crofton, K. M.; Thomas, R. S. Integrated Model of Chemical Perturbations of a Biological Pathway Using 18 in Vitro High-Throughput Screening Assays for the Estrogen Receptor. *Toxicol. Sci.* **2015**, *148*, 137−154.
(21) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.
(22) 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*, 2874−2884.
(23) Manganelli, S.; Roncaglioni, A.; Mansouri, K.; Judson, R. S.; Benfenati, E.; Manganaro, A.; Ruiz, P. Development, Validation and Integration of in Silico Models to Identify Androgen Active Chemicals. *Chemosphere* **2019**, *220*, 204−215.
(24) Mansouri, K.; Kleinstreuer, N.; Abdelaziz, A.; Alberga, D.; Alves, V. M.; Andersson, P. L.; Andrade, C. H.; Bai, F.; Balabin, I.; Ballabio, D.; Wedebye, E. B.; Benfenati, E.; Bhhatarai, B.; Boyer, S.; Chen, J.; Consonni, V.; Farag, S.; Fourches, D.; Garcìa-Sosa, A. T.; Gramatica, P.; Grisoni, F.; Grulke, C. M.; Hong, H.; Horvath, D.; Hu, X.; Huang, R.; Jeliazkova, N.; Li, J.; Li, X.; Liu, H.; Manganelli, S.; Mangiatordi, G.; Maran, U.; Marcou, G.; Martin, T.; Muratov, E.; Nguyen, D.; Nicolotti, O.; Nikolov, G. N.; Norinder, U.; Papa, E.; Petitjean, M.; Piir, G.; Poroikov, V.; Qiao, X.; Richard, A. M.; Roncaglioni, A.; Ruiz, P.; Rupakheti, C.; Sakkiah, S.; Sangion, A.; Schramm, K.; Selvaraj, C.; Shah, I.; Sild, S.; Sun, L.; Taboureau, O.; Tang, Y.; Tetko, I.; Todeschini, R.; Tong, W.; Trisciuzzi, D.; Tropsha, A.; Van Den Driessche, G.; Varnek, A.; Wang, Z.; Williams, A. J.; Xie, H.; Zakharov, A.; Zheng, Z.; Judson, R. S.; et al. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ. Health Perspect.* **2020**, *128*, No. 027002.
(25) Mathea, M.; Klingspohn, W.; Baumann, K. Chemoinformatic Classification Methods and Their Applicability Domain. *Mol. Inf.* **2016**, *35*, 160−180.
(26) Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate Comparison of Classification Performance Measures. *Chemom. Intell. Lab. Syst.* **2018**, *174*, 33−44.
(27) Marzo, M.; Kulkarni, S.; Manganaro, A.; Roncaglioni, A.; Wu, S.; Barton-Maclaren, T. S.; Lester, C.; Benfenati, E. Integrating in Silico Models to Enhance Predictivity for Developmental Toxicity. *Toxicology* **2016**, *370*, 127−137.
(28) Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure−Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867−878.
(29) Ballabio, D.; Todeschini, R.; Consonni, V. Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data. *Data Handl. Sci. Technol.* **2019**, *31*, 129−155.
(30) Ballabio, D.; Grisoni, F.; Consonni, V.; Todeschini, R. Integrated QSAR Models to Predict Acute Oral Systemic Toxicity. *Mol. Inf.* **2018**, *38*, No. 1800124.
(31) Fernández, A.; Lombardo, A.; Rallo, R.; Roncaglioni, A.; Giralt, F.; Benfenati, E. Quantitative Consensus of Bioaccumulation Models for Integrated Testing Strategies. *Environ. Int.* **2012**, *45*, 51−58.
(32) Borràs, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O. Data Fusion Methodologies for Food and Beverage Authentication and Quality Assessment − A Review. *Anal. Chim. Acta* **2015**, *891*, 1−14.
(33) Billoir, E.; Delignette-Muller, M. L.; Péry, A. R. R.; Charles, S. A Bayesian Approach to Analyzing Ecotoxicological Data. *Environ. Sci. Technol.* **2008**, *42*, 8978−8984.
(34) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(35) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, No. 20.
(36) Seber, G. A. F. *Multivariate Observations*; Wiley-Interscience, 2009.
(37) *Dragon (Software for Molecular Descriptor Calculation)*; Kode srl, 2017.
(38) *MATLAB R2018b*; The MathWorks, Inc.: Natick, MA, 2018.
(39) Ballabio, D. A MATLAB Toolbox for Principal Component Analysis and Unsupervised Exploration of Data Structure. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 1−9.
(40) Sailaukhanuly, Y.; Zhakupbekova, A.; Amutova, F.; Carlsen, L. On the Ranking of Chemicals Based on Their PBT Characteristics: Comparison of Different Ranking Methodologies Using Selected POPs as an Illustrative Example. *Chemosphere* **2013**, *90*, 112−117.
(41) Keeney, R. L.; Raiffa, H. *Decisions with Multiple Objectives*; Cambridge University Press, 1993.
(42) Hendriks, M. M. W. B.; de Boer, J. H.; Smilde, A. K.; Doornbos, D. A. Multicriteria Decision Making. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 175−191.
(43) Keller, H. R.; Massart, D. L.; Brans, J. P. Multicriteria Decision Making: A Case Study. *Chemom. Intell. Lab. Syst.* **1991**, *11*, 175−189.
(44) Vighi, M.; Barsi, A.; Focks, A.; Grisoni, F. Predictive Models in Ecotoxicology: Bridging the Gap between Scientific Progress and Regulatory Applicability—Remarks and Research Needs. *Integr. Environ. Assess. Manage.* **2019**, *15*, 345−351.

https://dx.doi.org/10.1021/acs.jcim.9b01057
J. Chem. Inf. Model. 2020, 60, 1215−1223