Journal of
Chemical Information
and Modeling

Article

pubs.acs.org/jcim

Applicability Domain Analysis (ADAN): A Robust Method for Assessing the Reliability of Drug Property Predictions

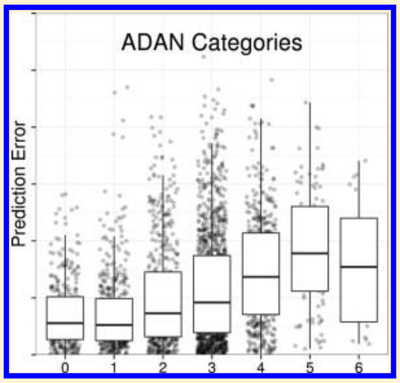
Pau Carrió,[‡] Marta Pinto,[§] Gerhard Ecker,[§] Ferran Sanz,[‡] and Manuel Pastor*,[‡]

[‡]Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, IMIM (Hospital del Mar Medical Research Institute), Dr. Aiguader, 88, E-08003 Barcelona, Spain

[§]Department of Medicinal Chemistry, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria

S *Supporting Information*

ABSTRACT: We report a novel method called ADAN (Applicability Domain ANalysis) for assessing the reliability of drug property predictions obtained by in silico methods. The assessment provided by ADAN is based on the comparison of the query compound with the training set, using six diverse similarity criteria. For every criterion, the query compound is considered out of range when the similarity value obtained is larger than the 95th percentile of the values obtained for the training set. The final outcome is a number in the range of 0−6 that expresses the number of unmet similarity criteria and allows classifying the query compound within seven reliability categories. Such categories can be further exploited to assign simpler reliability classes using a traffic light schema, to assign approximate confidence intervals or to mark the predictions as unreliable. The entire methodology has been validated simulating realistic conditions, where query compounds are structurally diverse from those in the training set. The validation exercise involved the construction of more than 1000 models. These models were built using a combination of training set, molecular descriptors, and modeling methods representative of the real predictive tasks performed in the eTOX project (a project whose objective is to predict in vivo toxicological end points in drug development). Validation results confirm the robustness of the proposed assessment methodology, which compares favorably with other classical methods based solely on the structural similarity of the compounds. ADAN characteristics make the method well-suited for estimate the quality of drug predictions obtained in extremely unfavorable conditions, like the prediction of drug toxicity end points.



INTRODUCTION

Evaluating the safety of new drug candidates is one of the critical steps in preclinical drug development. At a time when industry productivity is at an all-time low, while the cost and time required to develop a new drug are at an all-time high,[1] adverse effects must be identified early in the development process, to avoid spending money on inappropriate compounds.[2] Also, in early drug development, safety testing should require small amounts of product, avoid the use of animal experimentation,[3] and yield reliable results with negligible rates of false positive and negative results. Toxicologists are well-acquainted with these constraints and the struggle to develop suitable strategies that provide reasonable compromises.[4,5] Among these, computer-based (in silico) prediction technologies look promising (e.g., for prioritizing compounds, building an awareness of potential liabilities to look for, etc.). In silico predictions can produce fast results at any stage of drug development, consume no product, and require no animal testing. Indeed, in silico screening of some end points, like genotoxicity, skin sensitization, and hERG,[6] are widely used by the pharmaceutical industry. The extension of these methods for the prediction of organ and in vivo toxicity end points in drug development is one of the main objectives of project eTOX,[7] which is a public-private partnership project within the framework of the European Innovative Medicines Initiative (IMI).

Despite their many advantages, the adoption of in silico toxicity prediction methods and their acceptance by toxicologist and regulators depends critically on the quality of the results yielded. Predicted properties for a query compound must consistently and reliably reflect the true values. The use of wrong predictions (either false positives or false negatives[8]) might have severe negative consequences; therefore, the rate of these errors must be low. Ideally, the end users of in silico methods must have some guidance about the probability of obtaining wrong predictions and how far from the true values these predicted values will be. Without this information, the predictions are of very little practical utility and they cannot be used for making decisions involving large investments of time and money or, even worse, putting the health of patients at risk.

Evaluating the reliability of the predictions is an integral part of any method validation.[9] Typically, the method being evaluated is applied to a sample of compounds of known biological properties (validation sets), comparing the predicted and experimental values; the disagreements between both

Received: March 17, 2014
Published: May 1, 2014

values (error) can be expressed in different ways (e.g., standard deviation error of the predictions (SDEP), confidence intervals (CI), etc.) to represent the predictive quality of the method. When the method is applied to a new compound, it is tempting to use the results of the method validation directly to estimate the prediction error. However, this use of the results of this method validation would assume implicitly that the prediction for the query compound will be affected by an error similar to the average error obtained for the compounds in the validation set. This can be true when the query compound is very similar to those used for building or validating the model, but the assumption cannot be generalized. Models have certain Applicability Domain (AD) out of which the predictions are likely to be affected by larger errors.[10] There are different theoretical reasons that justify this model behavior. Statistical models (like the Quantitative Structure−Activity Relationships, QSAR) describe the association between the variation of some structural features and the variation of their biological properties within a collection of compounds (training set). If the query compound does not have the same structural scaffold, is missing variable structural features, or incorporates different ones, the predictions produced by the model could have no validity at all. Moreover, the biological properties of the query compound might involve biological mechanisms different from those described by the model and present in the training set compounds.

Classically, such AD problems were addressed by defining a discriminant function (compound in/out of the AD) that alerts when the predictions are likely to be affected by large errors. A detailed review of all the criteria used to define the AD is beyond the scope of this work, and interested readers can consult recent reviews, such as those presented in ref 11. However, in few words, the methods can be grouped in three large families: training set comparison, activity spaces, and model perturbation.

(a). Training Set Comparison. These methods are based on comparing the query compound with those of the training set. The characteristics being compared can be generic structural descriptors (e.g., fingerprints) or the specific properties used by the model for computing the predictions (molecular descriptors). The comparison can use Euclidean distance or other more-sophisticated metrics (e.g., Mahalanobis distance or city block distances),[11] and can be computed between the query compound and the centroid, the closest compound, or the k-nearest neighbors[12,13] of the training set. In some cases, instead of computing distances, the method simply ascertains that the query compound is within the range of the variables covered by the training set, either one by one or building a "convex hull" that encloses all compounds.[13] The criteria for defining a threshold that discriminates between compounds in and out are also diverse. In some cases, it is an adjustable parameter that changes from dataset to dataset and, in other cases, probabilistic methods are applied.

(b). Activity Spaces. These methods assume that the property space is not homogeneous, with respect to the quality of the predictions obtained. The expected reliability of the predictions is assessed by analyzing the quality of the predictions obtained in the vicinity of the query compound, using diverse approaches.[14,15]

(c). Model Perturbation. These methods are based in the analysis of the prediction stability after the introduction of perturbations, (e.g., by using different descriptors or by different bagged sets of data). A recent example can be found in ref 16.

Turning back to our original goal, obtaining an estimation of the prediction reliability, we must stress again that these AD estimators were designed to detect possible extrapolation errors but not to estimate the error for compounds within the AD. Consequently, AD estimators cannot be used directly to express how reliable is a prediction in a quantitative manner.[15] The strategy presented here combines the use of model validation results with methods used for AD assessment for estimating the reliability of single predictions. Classical model quality indexes, such as the model SDEP, will be considered as the "most optimistic" predictive reliability estimators. Then, it will be assumed that the quality of the predictions will suffer a constant decay, because the query compound is more dissimilar from the compounds in the validation set. Since this concept of dissimilarity is identical to the one introduced in the AD assessment, it could be quantified using the concepts and methods contributed by several authors in the literature mentioned above. Based on these ideas, we have developed ADAN (Applicability Domain ANalysis), a robust method for assessing the reliability of predictions. ADAN has been developed within the framework of eTOX to meet specific project needs and produce useful reliability indexes for a wide variety of prediction environments, including some that were classically considered highly unfavorable (e.g., complex end points, structurally unrelated query structures, etc.). However, the applicability of the method is not limited to the toxicology field and can be easily applied in other drug development contexts. The method design principles, the algorithm, and the results of an extensive validation exercise carried out on toxicologically relevant end points, will be reported here. The information provided would allow the method to be implemented in any programming language. Moreover, a complete implementation of ADAN as an R package can be downloaded, free of charge, from this web address: http://phi.imim.es/adan.

■ METHODS

The ADAN Method. The ADAN method was crafted to address specific needs of the eTOX project for assessing the reliability of in silico toxicity predictions. The eTOX project required that the reliability assessment was expressed in terms that were easy to understand, allowing end users to judge the validity of the outcome of predictive system for decision making (e.g., candidate prioritization). Robustness is also important and should be prioritized over the method accuracy or precision. Finally, the method must be able to work unsupervised, without the intervention of experts that adjust tunable parameters.

The basic design principle of the method is the comparison of the query compound with the training set. ADAN uses the results of model validation exercises (comparison of predicted versus experimental values for a validation sample) as the "most optimistic" estimators of the quality of any single prediction, and corrects these in function of the similarity between the query compound and the model. Another design principle is the assumption that a compound can be poorly predicted by a model because of several reasons. Hence, ADAN follows an eclectic approach and incorporates several binary criteria (in/out), some of which are classically used for the AD assessment. This approach represents a strategy to overcome the difficulties to describe molecular similarity in chemical spaces that are nonhomogeneously populated, often heavily clustered, by combining diverse similarity metrics. The number of criteria broken is then used to assign the query compound to a quality category, like in the widely used "Lipinski's rule of five" method for defining the drug likeness of a compound.[17] Compounds that break no rules (category 0) are likely to be very similar to the training set compounds and, therefore, likely to have prediction errors similar to those obtained in the validation exercise, while compounds that break an increasingly higher number of rule are likely to have larger differences and, therefore, are likely to suffer from larger errors. The criteria considered in ADAN are listed in Table 1, and the number of categories corresponds to the number of criteria considered plus one, labeled from 0 to 6.

Table 1. Criteria Used by ADAN

ID	description
A	distance between the query compound and the centroid of the training set
B	distance between the query compound and the closest compound in the training set
C	distance to model (DModX)
D	difference between the predicted biological property and the average biological property in the training set
E	difference between the predicted biological property and observed biological value for the closest compound in the training set
F	SDEP of the 5% of closest compounds in trainings set to the query compound

Distances mentioned in criteria A and B are calculated as Euclidean distances in the space of latent variables (LV) of a partial least squares (PLS) model built using the original molecular descriptors. The use of this projected space has the advantage over other methods (e.g., principal component analysis, PCA) of biasing the projection, giving more weight to the more biologically relevant features while removing redundancies due to the presence of highly correlated variables. In order to retain enough structural information, ADAN uses as many LV as were needed to explain at least 80% of the X variance. The Distance-to-Model (DModX)[18] used in criteria C is a model outlier diagnostic score that can be easily obtained from the X residuals of the projection on the aforementioned PLS model. Its interest resides in its ability to detect in the query compound the presence of characteristics not present in the training set and, hence, be able to distort the prediction results.

ADAN will compute two additional distances that make use of the compounds biological properties. Criterion D describes the differences between the predicted value for the query compound and the central value of the biological property on the training set. The rationale is that compounds for which we predicted extreme values are likely to have higher prediction errors. Criterion E is built by determining how different the predicted property for the query compound is, relative to the observed value for the closest compound in the property space.

Finally, ADAN incorporates an additional criterion (F) similar to SDEP-based criteria reported by other authors.[15] The distance measures the prediction accuracy of the 5% closest compounds to the query compound in the same LV space of metrics A and B. If the standard deviation of these errors is significantly higher than those obtained for the training set, the compound is considered to be located in a region where the predictions tend to be highly inaccurate and then is considered to be potentially unreliable.

All the above criteria yield a quantitative value that is converted to a binary output by comparing it with the 95% percentile of the training set. For example, a query compound is considered to break criteria A if the distance to the centroid measured for this compound is larger than 95% of the equivalent distances computed for the compounds of the training set.

The aggregation of the ADAN metrics is simple. This simplicity will make easier for the end users to understand why a model should not be used to predict a compound. In this sense, ADAN presents advantages with respect to complex aggregation schemes, such as the one proposed by Sheridan,[19] which are more difficult to interpret or tune.

For some uses, this number of criteria broken is informative enough. In particular, this scoring can be easily translated to a "traffic light" schema of coloring that can be applied in software graphic interfaces for expressing the expected reliability of a large number of predictions. For other uses, the same scoring allows one to define categories of compounds that can be assigned approximate confidence intervals (ACI), computed from the results of the model validation exercises. For example, if the SDEP obtained from model validation is 0.5, assuming an approximate normal distribution of the errors, we can build the 95% confidence interval around the predicted value as plus/minus 0.98 (1.96 times 0.5). Based on the assumption that compounds that break few ADAN rules are not too diverse from the validation set, we can set ACI equivalent to the model CI for these compounds. For compounds breaking more ADAN rules, the ACI can be widened by multiplying the CI by an appropriate coefficient. Finally, compounds breaking many rules cannot be expected to have reliable predictions and these values must be discarded. It must be stressed that this generic consideration does not allow one to identify the value of "few rules", "more rules", and "many rules", nor is the "appropriate coefficient" used to expand the ACI. The proposed mapping between the number of criteria broken and the ACI is based on an analysis of the ADAN method validation results that will be described later in this work. It must be stressed that these ACI are approximate and based on assumptions that might not be formally correct in all instances (e.g., normal distribution of the prediction errors). Still, approximate approaches are common in toxicology and the use of such rules is perfectly reasonable if the final results confirm the usefulness of the so-obtained results.

Summarizing, to assess the reliability of a prediction made for a query compound, the ADAN method uses, as input, the query compound molecular descriptors (x) and the predicted (y') value. The $\mathbf{x}$ vector is projected into an internal PLS model built ad hoc, obtaining the position of the query compound in the LV space, which are compared with those of the training set compounds for computing criteria A and B. The errors of the compounds closest to the query compound in this space are used to compute criteria F, and the X residuals of the PLS model are used to compute criteria C. The predicted y' value will be also compared with the activities of the training set to compute criteria D and E. The main output of the method is a single value in the range 0 to 6 that indicates the total number of ADAN criteria broken and defines the ADAN category of the query compound. This value can be used directly or combined with the results of the standard model validation for computing the prediction ACI.

Table 2. Characteristics of the Validation Datasets

dataset	end point	N^a	mean^b	SD^b	range^b	units	source
BZR	binding affinity to the benzodiazepine receptor	333	7.44	1.14	3.60−9.47	pIC_{50}	ref 20
COX-2	inhibition of cyclo-oxygenase-2	414	6.94	1.14	4.00−9.00	pIC_{50}	ref 20
DHFR	inhibition of dihydrofolate reductase	673	5.84	1.16	0.82−10.46	pIC_{50}	ref 20
ER	binding affinities to estrogen receptor	393	5.27	1.23	3.7−9.7	pIC_{50}	ref 20
hERG	binding affinities for the hERG potassium channel	356	5.65	1.23	2.36−9.05	pIC_{50}	ref 22
SOLU	aqueous solubility	1144	−3.06	2.10	−11.60−1.58	$\log_{10}$ (mol/L)	ref 21

[a]Number of compounds. [b]Mean, standard deviation, and range of the predicted property.

Validation Methodology. In order to validate the usefulness of the ADAN method, we applied ADAN in a large number of models. We made an effort to simulate models that can be considered representative of the conditions where it will be applied in practice. This principle guided the choice of the datasets, the training/validation set splitting method, and the modeling methodologies, as will be described in the following sections. It should be noted that these conditions are not optimal for obtaining good quality models. The purpose of this exercise is testing whether the ADAN method is able to discriminate between good from bad predictions in realistic situations and not to obtain accurate predictions. Indeed, the robustness of the method, understood as its ability to produce good assessments in highly unfavorable conditions was one of the key requirements, determinant for its usefulness in practice.

Datasets. The datasets used for this validation exercise (Table 2) are large and structurally diverse. The predicted properties are mainly binding to targets and antitargets used in early drug development. Six datasets were used; cyclo-oxygenase-2 (COX-2) inhibitors, benzodiazepine receptor (BZR) ligands, estrogen receptor (ER) ligands, dihydrofolate reductase (DHFR) inhibitors, potassium channel hERG blockers (hERG), and aqueous solubility values (SOLU). The datasets were extracted from the sources listed in Table 2.

The ER, COX-2, and BZR original datasets contained compounds with property values described with inequalities that were removed. COX-2, BZR, DHFR, and ER datasets contained already-curated three-dimensional (3D) structures provided by ref 20, so no extra curation was applied. The SMILES listed in the original sources for the SOLU[21] and hERG datasets[22] were converted to Mol files using OpenBabel 2.3.0.[23] All structures were protonated to pH 7.4 using Moka 1.1.0-RC3.[24,25] When necessary, 3D structures were generated using CORINA 2.4.[26,27]

Training-Test Set Splitting. The splitting algorithm used in this validation aims to simulate highly unfavorable conditions where the model is used to predict compounds clearly out of its AD. For every dataset, we obtained a PCA model of two principal components with the original molecular descriptors, such as the one represented in Figure 1. The variance explained by the model range between 39% and 98% (average of 73%). In such space, the compounds tend to spread according to generic molecular properties, such as molecular size and polarity (see the molecular weight represented in Figure 1), thus producing an efficient separation of congeneric series in separate clusters. The size of the training set is fixed to 35% of the compounds in the original dataset. A "seed compound" is picked randomly and compounds around it are picked randomly but with probability of being chosen that grows exponentially for the compounds closer to the seed compound in the property scores space. The effect of this algorithm can be seen in Figure 1, where the compounds in the training set are represented as crosses. In this validation, every predictive model is built using a training set selected as described, while the rest of the compounds (65%) are used as an external prediction set. Therefore, the method was designed to minimize the probabilities that the predicted compounds belong to the same congeneric series of the compounds in the training set. The entire procedure is repeated 31 times for every dataset, thus building 31 different models with 31 different training and test sets, in order to minimize the effect of very favorable or unfavorable training set choices in the final results.

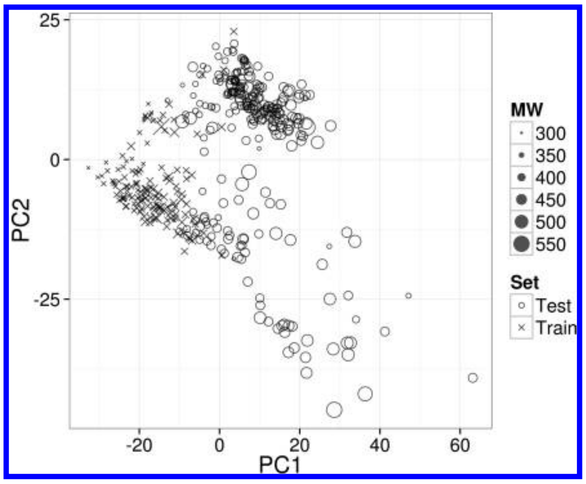


Figure 1. PCA scores space for the COX-2-ADRIANA dataset used to select the training set (crosses) as the 35% of compounds closest to a randomly selected seed compound. The molecular weight (MW) was represented in the size of the symbols to illustrate the structural diversity of the compounds.

Model Building. For all the datasets, we computed three types of molecular descriptors commonly used in QSAR modeling: PaDEL, ADRIANA, and PENTACLE. PaDEL descriptors[28] were generated using the free software PaDEL-Descriptor. ADRIANA descriptors[29] were generated with ADRIANA.Code.[30] PaDEL and ADRIANA contain a mixture of 2D and 3D descriptors. Finally, PENTACLE generates 3D GRid INdependent Descriptors (GRIND) that have been used extensively in many fields of drug development.[31,32] The PaDEL and ADRIANA descriptors were centered and scaled to unit variance, while the PENTACLE descriptors were only centered, as recommended by the authors.[31]

The models were built using two different modeling methods: Partial Least Squares (PLS) and Random Forest (RF), as implemented in the R packages pls[33] and randomForest,[34] respectively. In the case of PaDEL and ADRIANA descriptors, we performed a feature selection based on RF. Briefly, the prediction error on the out-of-bag portion of the data is recorded for each tree. Features with a mean decrease in accuracy higher than the 95th percentile were selected[35] to build the prediction model. The dimensionality of the PLS models was decided for each model, using the results of a Leave-One-Out (LOO) cross-validation: first, we build a model with 50 LVs, but the final number of LV used in the prediction model corresponds to the lowest model dimensionality producing a q^2 above the 90th percentile.

Models were built for all the possible combination of the 2 modeling methods, 3 molecular descriptors, 6 datasets, and 31 splits, producing a gross amount of 1116 models. Because of the highly unfavorable training/test set splitting, in many cases, we failed to obtain models of acceptable quality. Figure 2 depicts the predictive quality of the models obtained in terms of LOO cross-validation q^2.

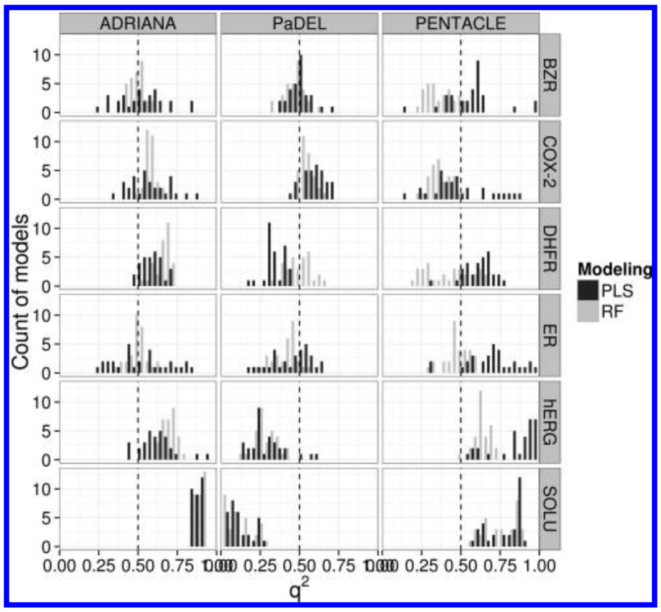


Figure 2. Distribution of the LOO cross-validation results (q^2) obtained for diverse modeling settings (modeling method, molecular descriptors, and datasets). Only the models with $q^2 > 0.5$ (dashed line) were selected for the validation exercise.

For the purposes of ADAN validation, we selected only the models with a q^2 value of >0.5 (643 out of 1116, 58%).

Structural Similarity-Based Applicability Domain Assessment. In order to obtain data for comparing our method with the current state-of-the-art, we computed generic structural similarity criteria based on MACCS fingerprints and Tanimoto distances for all the series. For every prediction, the AD was assessed quantifying the structural similarity between the query compound and the closest compound in the training set of this particular model. Tanimoto distances were computed from the MACCS fingerprints obtained with R package rcdk, based on CDK.[36] The Tanimoto distances were represented as bins of 1-Tanimoto to obtain a distance metric easier to compare with ADAN categories: compounds most similar to the training set are located on the left-hand side and those most dissimilar to the training set are located on the right-hand side.

ANOVA Analysis. The absolute values of the predictions errors obtained for all the models passing the $q^2 = 0.5$ filter were compiled in a single table. The net effect and the statistical significance of different factors in the magnitude of these errors was analyzed using one-way ANOVA for a single modeling setting and multiway ANOVA for all the datasets. In the first case, we included as a single factor the ADAN category, which takes seven values. In the second case, the following factors were included (number of values given in parentheses): datasets (6), molecular descriptors (3), modeling methods (2), ADAN categories (7), and randomization runs (31). For the computation, we used the aov command of the R package stats. In addition, the values of the model coefficients for each factor, always relative to the first value of the corresponding factor, were computed, as well as their statistical significance (p-value).

RESULTS

The ADAN method (see Methods section) was validated on a collection of 643 models built specifically to represent commonly found situations in the toxicology field. Before entering into the analysis of the ADAN performance for quantifying the reliability of the predictions, the characteristics of the models will be described briefly.

Quality of the Models. The best models were obtained for the SOLU dataset (average $q^2 = 0.83$) and the worst for the BZR and COX-2 datasets (average $q^2 = 0.59$ and 0.60, respectively). This is justified, in part, by the size of the datasets: 1144 compounds for SOLU versus 333 compounds for BZR and 414 compounds for COX-2. Also, the solubility end point represents a purely physicochemical property, while the BZR or COX-2 binding represents a far more complex phenomenon. Not less important is the fact that the experimental procedure is more complex and the experimental results (compiled from heterogeneous sources) can be expected to have larger errors. Table 3 also shows that, generally, PaDEL descriptors performed worse than other descriptors and for one dataset (SOLU), they even failed to produce any model that passes the filtering criteria.

Table 3. Average Predictive Quality of the Models (PLS and RF) Used for the ADAN Validation (LOO Cross-Validation q^2) Grouped by Datasets and Molecular Descriptors

	BZR	COX-2	DHFR	ER	hERG	SOLU	
ADRIANA	0.58	0.60	0.63	0.61	0.67	0.88	0.68
PaDEL	0.55	0.58	0.55	0.56	0.58	na	0.57
PENTACLE	0.63	0.70	0.63	0.68	0.75	0.77	0.71
	0.59	0.60	0.62	0.64	0.70	0.83	

With respect to the modeling methods, the average q^2 values for PLS and RF were 0.69 and 0.65, respectively (data not shown in Table 3), indicating that the PLS method performed slightly better than the RF method. A more-detailed description of the model performance and prediction error is provided as Supporting Information (SI).

Performance of ADAN. The prediction errors obtained for all compounds in the all the validations models were compiled and analyzed with the main objective of confirming the validity of the ADAN method. If the method works correctly, the errors will be consistently large for higher ADAN categories (compounds breaking more ADAN criteria). In addition, the impact of other factors on these errors (datasets, molecular descriptors, and modeling method) will be also analyzed.

Before presenting the results of the entire analysis, it is worth focusing on a single model setting in order to understand the tools used here to report the model errors. Figure 3 represents the prediction errors obtained by the models built using the COX-2 dataset, PENTACLE molecular descriptors, and the PLS method (hereinafter called COX-2-PENTACLE-PLS).

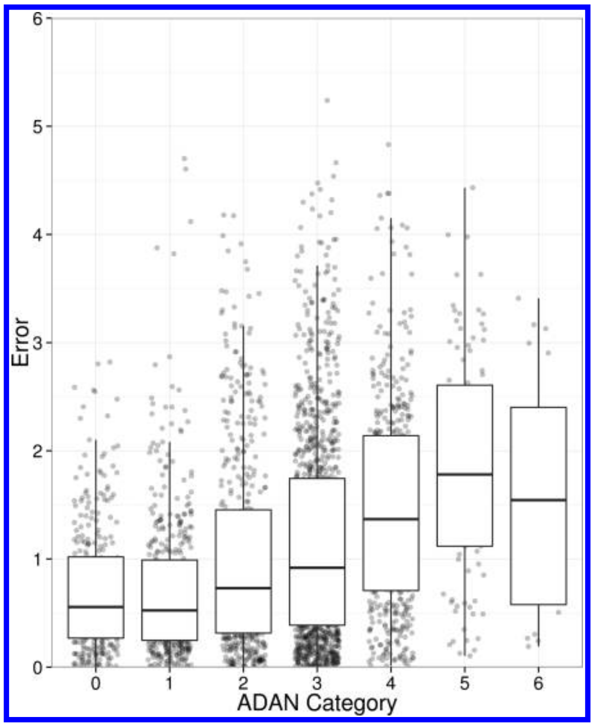


Figure 3. Prediction errors obtained for models generated using the COX-2 dataset, PENTACLE molecular descriptors, and the PLS modeling method. The errors were grouped according to the ADAN categories assigned to the predicted compounds.

The errors were grouped according to the ADAN categories to which the compounds belong and were depicted as standard box plots: the thick central line represents the median, the box delimits the interquartilic range (IQR, defined as the distance between the first and third quartiles), the upper whisker extends from the box top to the highest value, excluding outliers with values larger than 2 IQRs and the lower whisker extends from the box bottom to the lowest value, also excluding similar outliers. Data beyond the end of the whiskers are outliers and plotted as points. The actual error values were also plotted in the background.

As we can see, the prediction errors obtained for compounds in the lower ADAN categories are smaller and have smaller dispersion than those obtained for compounds in the higher categories. This was exactly the expected result and confirms the ability of the ADAN method to classify the predictions according to their reliability. The one-way ANOVA analysis of these results further confirm the statistical significance of the overall observed differences ($p < 2.10^{-16}$), as well as the significance of the coefficient estimates for each class (most significant p-value is 0.0001 for category 4). The number of compounds within each ADAN category is variable, and for this modeling setting, the population in classes 5 and 6 is relatively small, even if the distribution is variable and largely depends on the characteristics of the datasets. Also, it is worth reminding that the number of predictions is dependent on the number of models that pass the aforementioned filtering procedure; for some modeling settings (datasets, molecular descriptors, and modeling methods), many models passed the filtering, whereas, in some other cases, we obtained only a few models. An extreme example is the case of PaDEL-SOLU, for which no single model passed the filtering (see Figure 2A).

The same analysis was extended to the models obtained using all the modeling settings (datasets, molecular descriptors, and modeling methods). Figure 4 summarizes the values of the prediction errors obtained for the diverse ADAN categories using all modeling settings. For PLS models (Figure 4A), the trend observed previously for COX-2-PENTACLE-PLS was rather general. Consistently, the median errors obtained for compounds belonging to the highest ADAN categories are larger than the errors obtained for the lowest ones. This is

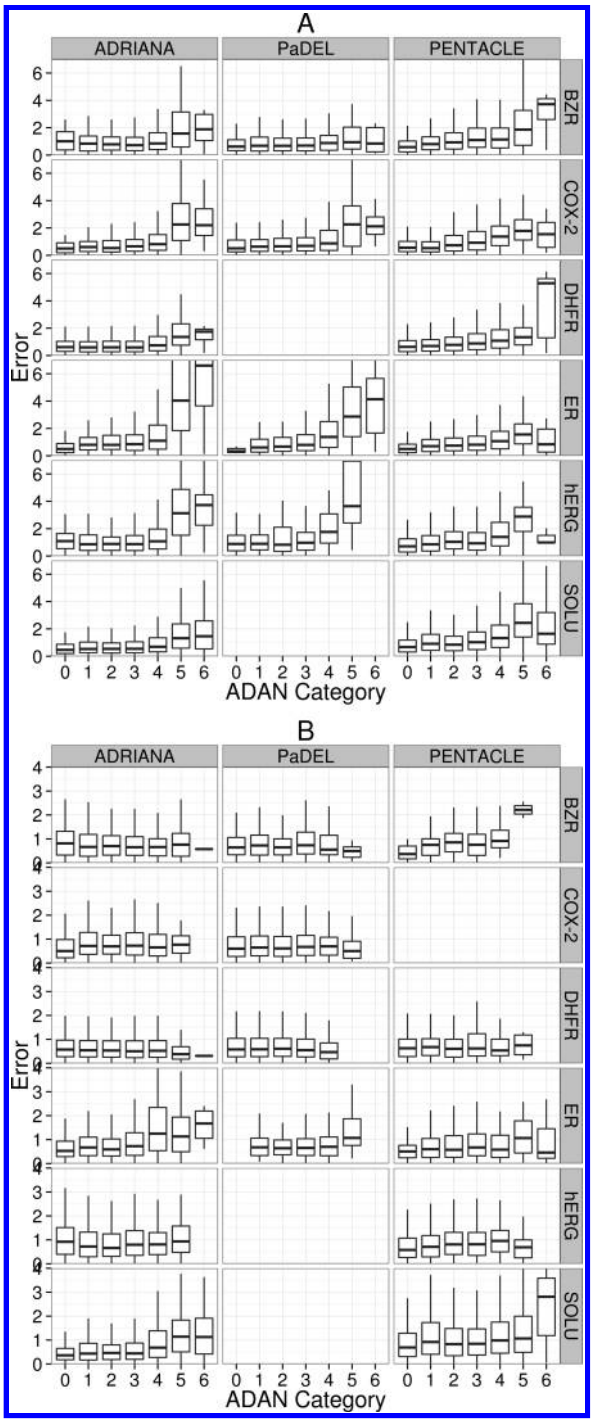


Figure 4. Boxplots representing the prediction errors obtained for diverse modeling settings, grouped by the ADAN category of the predicted compounds: (A) PLS results and (B) RF results.

observed for all datasets and molecular descriptors, even if there are some deviations that are worthy of comment. The worst results are obtained for PaDEL models, which are much worse (average q^2 value of 0.57) than models obtained using other molecular descriptors (average q^2 values of 0.68 and 0.71 for ADRIANA and PENTACLE, respectively), thus evidencing severe model limitations not associated with the individual properties of the compounds, already described in Table 3. Another observation is the clear deviation observed for category 6 in some models obtained with PENTACLE descriptors (ER, hERG, and SOLU datasets). This effect was due to the aforementioned population differences observed for higher classes and has no major relevance.

For models obtained using Random Forest (Figure 4B), the trend is still present but less evident. This is partially explained by the fact that, in this modeling method, the predictions are a weighted average of values present in the training set. Therefore, prediction errors in RF cannot be larger than the differences between two activity values of the training set and do not grow indefinitely, as is the case for PLS. This phenomenon is general and can be appreciated in the scale of the error plots—higher for PLS than for RF—but it is more evident in series such as COX-2, DHFR, and hERG, where the distribution of the activity values is very narrow and symmetrical (see Figure SI1 in the Supporting Information), than in series such as BZR, SOLU, and ER. As a consequence, for RF models, the correlation between the ADAN classes and the absolute value of the prediction error is only evident for series characterized by having an acceptable distribution of the activity values (such as BZ, SOLU, and ER), in particular for the modeling settings producing models of better quality (e.g., ADRIANA-SOLU) where the prediction error is very small for compounds in the lowest ADAN classes. Conversely, series where the activity values are not well-distributed, especially for modeling settings that yield low-quality models (e.g., PaDEL-DHFR), produce no good predictions (no low errors), and for the bad predictions, the errors have similar sizes in all instances. As a consequence, no growing trend can be observed in Figure 4B.

The statistical significance of the differences observed in these prediction errors for different ADAN classes can be assessed using multiway ANOVA. In our analysis, the absolute values of the prediction errors were used as dependent variables and we included the ADAN categories, the modeling settings (datasets, molecular descriptor, modeling methods), and the training/test splitting randomization run as factors, as described in the Methods section. The ANOVA table (Table 4) shows that all these factors, including the ADAN category, are highly significant, with p-values of <0.0001 in all cases.

A more-detailed analysis of the ANOVA results shows that all coefficients (Table 5) representing the effect of the individual values of the ADAN category variable are also significant at 99% confidence level and show roughly growing values, further supporting their value as estimators of the prediction error. Interestingly, the large values obtained for ADAN categories 2 and 5 can represent frontier categories between compounds with low and high prediction errors. This observation will help us to define alternative thresholds when a smaller number of categories is needed.

The observed ability of ADAN to describe the prediction reliability in our validation sets compares favorably with other classical assessment methods. For example, Figure 5A and 5B represents the structural similarity of the predicted compounds with the training data (quantified in terms of Tanimoto distances, see the Methods section) and the prediction error, in a way that allow a direct comparison with the plots in Figure 4. These results confirm the general belief that structural similarity with the training set compounds is an important component of any prediction reliability assessment, and, in most cases, the errors obtained for the largest structural similarity classes are largest than those obtained for more-similar compounds. However, the trend is unstable and behaves very differently for diverse series and modeling settings, making it impossible to determine a generally valid cutoff value that can be used to decide about the reliability of the predictions. For example, a value of 1-Tanimoto of <0.5 can be useful to recognize good predictions in ER-ADRIANA-PLS, but the same cutoff has no value in ER-PENTACLE-PLS or SOLU-PENTACLE-PLS (Figure 5A) therefore making this metric unsuitable for unsupervised assessment.

At this point, it is worth remembering that the ADAN category represents only the number of criteria broken. The contribution of every individual metric to this figure is shown in Figure 6. The frequency with which every individual criterion is broken in the diverse modeling settings is represented by the symbol size. The results show remarkable similarities across the diverse datasets and model settings, but there are also differences that account for the ability of the method to adapt to the characteristics of the diverse models. Generally, category 1 represents compounds breaking metric C (the distance to model, DModX), which is the most sensitive in all cases. Categories 2 and 3 include compounds that also break criteria A and B (the distance between the query compound and training set centroid or to the closest compound in the X space, respectively). Criteria D and E (difference between the predicted value and the average y-value, and the y difference with the closest compound of the training datasets, respectively) appear next in importance and are often coupled.

Table 4. Results of the Analysis of Variance (ANOVA), Describing the Association between the Prediction Errors, the ADAN Categories of the Compounds, and the Modeling Settings (Datasets, Molecular Descriptors, Modeling Method, and Randomization Run)

	Df[a]	Sum Sq[b]	Mean Sq[c]	F-value	p-value
dataset	5	5920	1184	392	<0.0001
molecular descriptor	2	100	50	16	<0.0001
regression method	1	2968	2968	983	<0.0001
ADAN category	6	38 059	6343	2101	<0.0001
run	30	1767	59	19	<0.0001
residuals	2 495 556	727 294	3		

[a] Degrees of freedom. [b] Sum of squares. [c] Mean squares.

Table 5. Coefficients of the ANOVA Linear Model for the ADAN Categories and Their Significance

ADAN category	coefficient estimate	p-value
0	ref. value	na
1	0.096	<0.0001
2	0.122	<0.0001
3	0.151	<0.0001
4	0.413	<0.0001
5	1.953	<0.0001
6	3.435	<0.0001

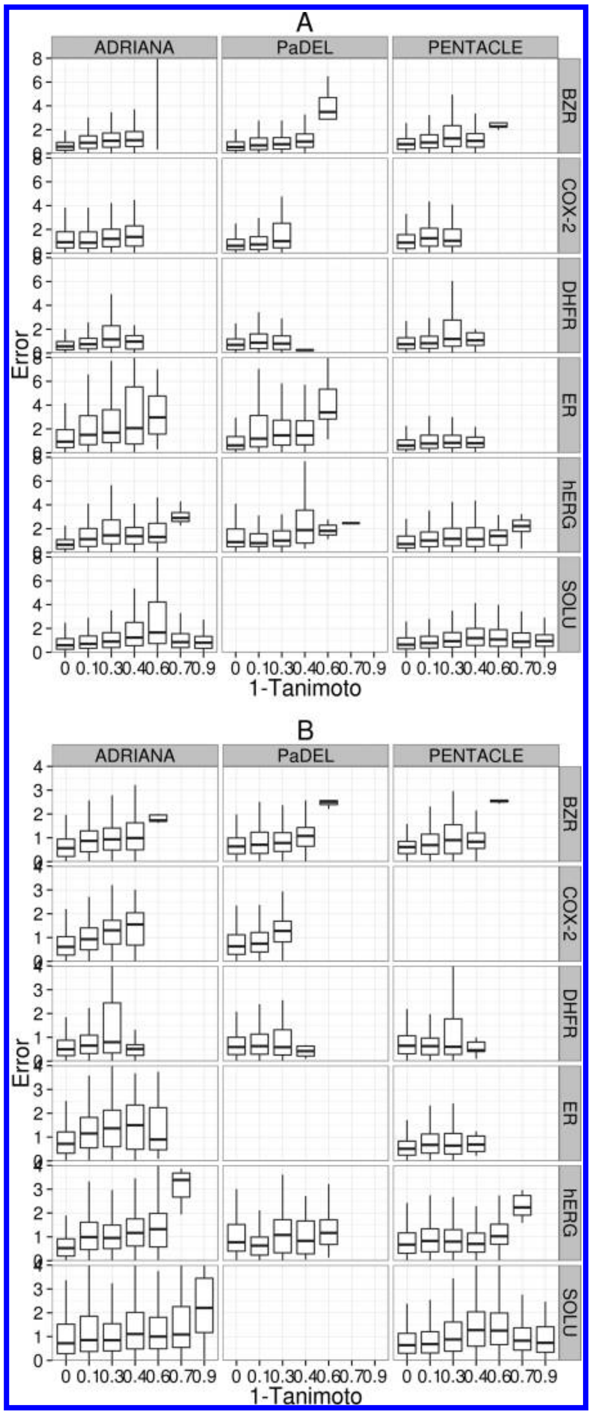


Figure 5. Boxplots representing the errors of the predictions obtained with (A) PLS and (B) RF plotted against the structural similarity of the predicted compounds with the training set. This similarity was quantified using the Tanimoto distances between the query compounds and the closest compound in the training datasets (see Methods section) and is represented as 1-Tanimoto, so the lowest values (on the left-hand side) correspond with the most similar compounds and the highest values (on the right-hand side) correspond to the most different compounds. Data points were binned in seven groups as provided with ADAN.

On the other side, criterion F (distance in Y-space) is the less-sensitive metric and is broken mainly in compounds that have already broken most of the criteria, (categories 5 and 6) with few exceptions. These results confirm the existence of diverse potential sources of error, not equally populated among the compounds and further stress the importance of considering multiple criteria for a proper assessment of the prediction reliability. The covariance of the different ADAN criteria was also investigated by computing the matrix of phi coefficients between all the criteria pairs (as described in the Supporting Information), for the studied modeling settings. The results (shown in Figures S8 and S9 in the Supporting Information) confirm that the ADAN criteria are rather independent. Only criteria A-B and, to a lesser extent, criteria D-E exhibit some consistent covariance for some modeling settings, even if these are much weaker for other settings. These results further confirm that the diverse criteria constitute independent, nonredundant sources of information that can be combined for the assessment of the prediction reliability.

As proposed in the Methods section, the ADAN categories can be used to compute approximate confidence intervals (ACIs) for each class; the SDEP values obtained in the regular model validation are used for compounds breaking few rules, widening the CI by a certain factor that depends on the number of criteria broken. As a first step for validating this idea, we assumed that the errors follow a normal distribution and estimated 95% CI of the predictions using the standard statistical method (adding and subtracting 1.96 times the model SDEP). We then tested the validity of such CI using our validation set, calculating how frequently the true (experimental) value is within the CI for diverse modeling settings. According to the CI definition, if the CI were independent of the ADAN category, the true value should be enclosed by the CI for ~95% of the predictions and out of it for the other 5%. Figure 7 shows that the actual proportion of true values within the CI is lower, even for the compounds in category 0, which is something that is justified by the extreme validation conditions that break the assumptions under which these CIs can behave as expected. Importantly, Figure 7 also shows that the proportion of true values within the CI decreases consistently for higher ADAN categories. The regular CI could be valid for the compounds belonging to the lowest ADAN categories, but for higher categories, the same CI must be expanded by a certain factor in order to capture the proportion of true values described by the confidence level.

Assigning a value to the aforementioned "expansion factor" is not trivial. The diverse simplifications introduced in the method do not allow a formal assignment, based on theoretical considerations. Moreover, the results shown in Figure 7 suggest that the proportion of true values in/out of the CI will be rather diverse for different datasets. As a pragmatic approach for the definition of ACI, we propose using SDEP-derived CI, similar to that used previously, for compounds in categories 0 and 1 ("green type", G) and to double the CI width for compounds in categories 2 and 3 ("yellow type", Y). Compounds in categories 4 and higher ("red type", R) also should be considered out of the AD for estimating quantitatively the reliability of their predicted properties and we recommend discarding them as "not reliable". The result of applying this approach in our validation models is illustrated in Figure 8. As we can see, the adoption of such ACIs for Y compounds allows one to capture, within the ACI range, a large proportion of the true values, often higher than the proportion present for G compounds. The proportion for R compound is also shown in the graphic, even if we strongly recommend avoiding their use for practical purposes. In order to understand Figure 8 properly, we must point out that the ACIs can be very

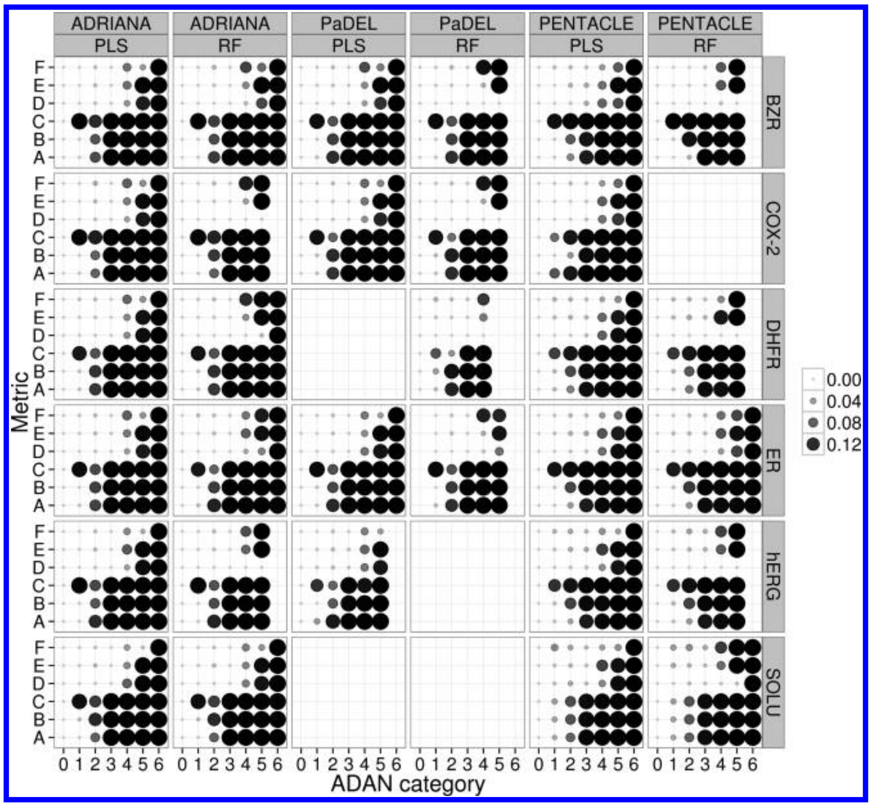


Figure 6. Contribution of the ADAN criteria to define the ADAN category. The columns represent the datasets used, and the rows represent the type of molecular descriptor−modeling method.

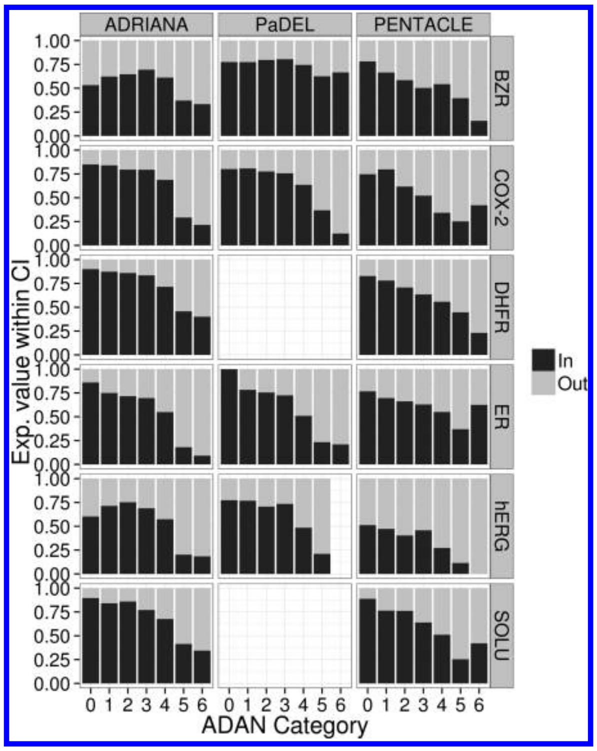


Figure 7. Proportion of predictions for which the true (experimental) values are enclosed within the CI (built as predictions plus/minus 1.96 times SDEP), for different ADAN categories and modeling settings (only PLS models, diverse datasets, and molecular descriptors). The black and gray sections of the columns represent the median proportion of predictions for which the true value is in and out of the CI, respectively.

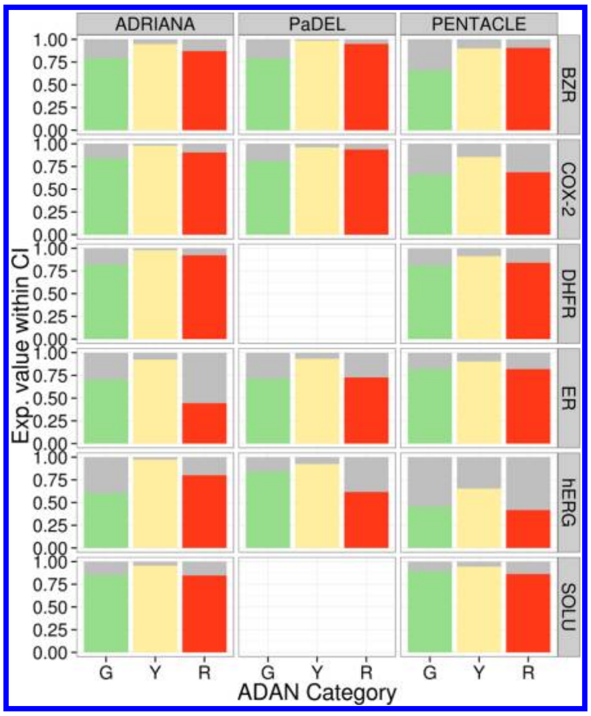


Figure 8. Proportion of predictions for which the true (experimental) values are enclosed within the ACI (see text for definition), for three categories of compounds (G, ADAN categories 0 and 1; Y, ADAN categories 2 and 3; and R, ADAN categories 4 and higher), for diverse modeling settings (PLS models, for diverse datasets and molecular descriptors). The colored and gray sections of the columns represent the median proportion of compounds in and out of the ACI, respectively.

wide for the worst models (with large SDEP); this produces an artificially large proportion of true values within the ACI, even if their practical usefulness is limited.

DISCUSSION

One of the principles that guided the design of ADAN was the belief that approximated but comprehensive descriptions of the phenomena are better than accurate models with a very narrow application field. The former approaches tend to yield generally useful results, while the latter approaches produce excellent results, but only seldom. Translating this principle to the practice implied identifying potential sources of prediction error, common to diverse computational methods and combining them into a sensible score, using a simple and general equation that is not dependent on a constellation of tunable and adjustable parameters. The ADAN method incorporates criteria that describe the distance of the query compound to the model (A, B), local predictive accuracy (F), outlier checking (C), and distance of the predicted biological property (D, E). These were combined by computing a simple count of criteria broken, based on our observation that the compounds that accumulate higher errors tend to be so because of many different reasons at the same time (e.g., being model outliers, far from the centroid of the training set and also far from any single compound). The strategy chosen for avoiding tunable cutoffs was to contrast the metrics obtained for every criteria with the distribution of the same metric in the model training set. The 95% percentile was found to be a reasonable criterion for discriminating, in a general way, values commonly found in the training set from those found only exceptionally.

The actual result of the ADAN method is a single score from 0 to 6. Depending on the intended use, this value can be translated to other scales. In the Results section, we have shown how this scoring can be used to define three categories (G/Y/R) for which ACI can be computed. Other uses and scales are possible. All in all, it must be stressed that, even if most of the criteria embedded into ADAN were designed only for a yes/no discrimination, by combining them into the ADAN score, we obtain a richer reliability assessment, which allows for a far more flexible application.

The results of the validation reported here show a clear and consistent relationship between the actual prediction error and the ADAN classes for all series and molecular descriptor tests, in the case of PLS models. For RF models, the relationship was clear for the series with best models (SOLU) but was not apparent for the worst modeling settings or those with very narrow distributions of the biological properties. This effect, as was described in the results, is partially an artifact that is due to the peculiarities of RF and should not affect the performance of the methods in situations less extreme than those used in this validation exercise.

An aspect of the ADAN methodology that deserves further discussion is the choice of the PLS projected space for computing distances in criteria A, B, and F. There has been wide discussion in the literature about the most convenient metric for computing molecular similarity.[37] In our opinion, a good metric must prioritize biologically relevant characteristics of the compounds. Only in this metric can two close compounds be expected to have similar biological properties, avoiding activity cliffs.[38,39] Unfortunately, the selection of biologically relevant properties is not simple, particularly if we want to delegate this task to unsupervised computational methods. The approach followed in ADAN was to build an ad hoc PLS model that correlates the original molecular descriptors with the end point, generating a space of latent variables[18,40] representing a biased metric in which the variables most correlated with the biological end point are assigned a higher weight. Interestingly, the effect of other properties (less correlated with the biological end point) is not completely neglected, something that can be useful to incorporate, at least partially, features not present in the training set, but interesting for predicting structurally diverse compounds. Indeed, the choice of this metric was the result of preliminary analysis (not shown here) in which the PLS-LV performed consistently better than metrics based on Euclidean distances or PCA−PC scores.

The aforementioned vocation for obtaining a robust and widely applicable method is also reflected in the design of the method validation. Instead of selecting a few well-studied datasets and a mild training/test set splitting algorithm, we decided to use a wide range of datasets, representing challenging end points, and produce models that combine diverse molecular descriptors and modeling methods. This collection is representative of the predictive models used in the project eTOX, applied in the most extreme conditions we could imagine, so the performance of the ADAN method in practice can be realistically represented by the results reported here.

CONCLUSIONS

Here, we have reported a novel general methodology for estimating the reliability of predictions obtained by computational methods. The strategy applied intends to incorporate the main sources of error and provide robust approximate estimators of the prediction errors. The method contains no tunable parameters and can work unsupervised, producing reasonable estimations of the prediction errors in highly unfavorable conditions.

Even if several other methods for carrying out a similar assessment have been published, we found no suitable method covering the need of the eTOX project. With respect to other methods, ADAN has the following features:

- Combines into a single score multiple criteria representing diverse source of prediction error;
- Avoids the use of arbitrary cutoff values or tunable parameters; and
- Provides an integrated and understandable measurement of the prediction reliability.

The results of the strict validation exercise reported here show that the ADAN method provides stable and robust assessment of the prediction error for a wide variety of datasets, modeling methods, and molecular descriptors. The assessment obtained, even if approximate, has been demonstrated to be statistically significant and compares favorably with classical methods based on single criteria (structural similarity). Moreover, the method yields results that can be easily translated to qualitative scales (e.g., traffic lights based) or used to compute approximate expanded confidence intervals. In this sense, the method is far more useful than classical applicability domain methods that yield only a binary (yes/no) result.

The ADAN method is simple to implement and can be easily reproduced or adapted using the description Figure S1 therein. However, we provide a free implementation of the method using open source software R.

The principles of the ADAN methodology described here can be easily applied for obtaining prediction reliability assessment in other computational methodologies. For example, criteria A, B and C can be directly used for assigning simpler reliability categories to predictions produced by models built using qualitative end points. The predictions obtained with non statistical methods (e.g., pharmacophoric models) can also be characterized analyzing similarities between the data sets used to develop the pharmacophore (conceptually similar to the training set) and the query compounds. At present, we are working to extend the ADAN methodology to these fields and to validate its usefulness.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Activity values centered to the mean (Figure S1); compounds inside the defined CI for RF modeling (Figures S2 and S3); count of compounds per ADAN category with PLS modeling (Figure S4); count of compounds per ADAN category with RF modeling (Figure S5); prediction error for PLS and RF (Figure S6); ADAN and prediction error example (Figure S7); phi coefficients between ADAN criteria pairs for all studied series (PLS modeling) (Figure S8); phi coefficients between ADAN criteria pairs for all studied series (RF modeling) (Figure S9). Average predictive quality of the PLS models used for the ADAN validation grouped by datasets and molecular descriptors (Table S1). Average predictive quality of the RF models used for the ADAN validation grouped by datasets and molecular descriptors (Table S2). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: manuel.pastor@upf.edu.

Author Contributions

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

Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking, under Grant Agreement No. 115002 (eTOX), resources of which are composed of a financial contribution from the European Union's Seventh Framework Programme (FP7/2007−2013) and EFPIA companies' in kind contributions.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

ACI, approximate confidence interval; AD, applicability domain; ADAN, applicability domain analysis; ANOVA, analysis of variance; BZR, benzodiazepine receptor ligands; CI, confidence interval; COX-2, Cyclooxygenase-2; DHFR dihydrofolate reductase; DmodX, distance to model; ER, estrogen receptor; hERG, human ether-a-go-go related gene; LOO, leave one out; PC, principal component; PCA, principal component analysis; PLS, partial least-squares; QSAR, quantitative structure−activity relationships; SDEP, standard deviation error of the predictions; SOLU, solubility

■ REFERENCES

(1) Stevens, J. L.; Baker, T. K. The Future of Drug Safety Testing: Expanding the View and Narrowing the Focus. *Drug Discovery Today* **2009**, *14*, 162−167.

(2) Modi, S.; Hughes, M.; Garrow, A.; White, A. The Value of in Silico Chemistry in the Safety Assessment of Chemicals in the Consumer Goods and Pharmaceutical Industries. *Drug Discovery Today* **2012**, *17*, 135−142.

(3) REACH. *European Community Regulation on chemicals and their safe use*, http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed Sept. 9, 2013).

(4) Meanwell, N. A. Improving Drug Candidates by Design: A Focus on Physicochemical Properties as a Means of Improving Compound Disposition and Safety. *Chem. Res. Toxicol.* **2011**, *24*, 1420−1456.

(5) Bass, A. S.; Cartwright, M. E.; Mahon, C.; Morrison, R.; Snyder, R.; McNamara, P.; Bradley, P.; Zhou, Y.-Y.; Hunter, J. Exploratory Drug Safety: a Discovery Strategy to Reduce Attrition in Development. *J. Pharmacol. Toxicol. Methods* **2009**, *60*, 69−78.

(6) Car, B. Enabling Technologies in Reducing Drug Attrition Due to Safety Failures. *Am. Drug Discovery* **2006**, *1*, 53−56.

(7) Briggs, K.; Cases, M.; Heard, D. J.; Pastor, M.; Pognan, F.; Sanz, F.; Schwab, C. H.; Steger-Hartmann, T.; Sutter, A.; Watson, D. K.; Wichard, J. D. Inroads to Predict in Vivo Toxicology—An Introduction to the eTOX Project. *Int. J. Mol. Sci.* **2012**, *13*, 3820−3846.

(8) Tetko, I. V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D. C.; Poda, G. I. Can We Estimate the Accuracy of ADME-Tox Predictions? *Drug Discovery Today* **2006**, *11*, 700−707.

(9) Weaver, S.; Gleeson, M. P. The Importance of the Domain of Applicability in QSAR Modeling. *J. Mol. Graph. Model.* **2008**, *26*, 1315−1326.

(10) Netzeva, T.; Worth, A.; et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure−Activity Relationships. *Altern. Lab. Anim.* **2005**, *33*, 155−173.

(11) 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.

(12) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel k-Nearest Neighbours Approach to Assess the Applicability Domain of a QSAR Model for Reliable Predictions. *J. Cheminform.* **2013**, *5*, 27−36.

(13) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA, Altern. Lab. Anim.* **2005**, *33*, 445−459.

(14) Keefer, C.; Kauffman, G.; Gupta, R. Interpretable, Probability-Based Confidence Metric for Continuous Quantitative Structure−Activity Relationship Models. *J. Chem. Inf. Model.* **2013**, *53*, 368−383.

(15) Briesemeister, S.; Rahnenführer, J.; Kohlbacher, O. No Longer Confidential: Estimating the Confidence of Individual Regression Predictions. *PLoS One* **2012**, *7*, e48723.

(16) Sheridan, R. P. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J. Chem. Inf. Model.* **2012**, *52*, 814−823.

(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(18) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361−1375.

(19) Sheridan, R. P. Using Random Forest to Model the Domain Applicability of Another Random Forest Model. *J. Chem. Inf. Model.* **2013**, *53*, 2837−2850.

(20) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification

Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.

(21) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.

(22) Obiol-Pardo, C.; Gomis-Tena, J.; Sanz, F.; Saiz, J.; Pastor, M. A Multiscale Simulation System for the Prediction of Drug-Induced Cardiotoxicity. *J. Chem. Inf. Model.* **2011**, *51*, 483−492.

(23) O'Boyle, N. M.; Banck, M.; James, C. A; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33−47.

(24) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **2009**, *49*, 68−75.

(25) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pK_a Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172−2181.

(26) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(27) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(28) Yap, C. W. E. I. Software News and Update PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2010**, *32*, 1466−1474.

(29) Gasteiger, J. Of Molecules and Humans. *J. Med. Chem.* **2006**, *49*, 6429−6434.

(30) *ADRIANA.Code*; Molecular Networks GmbH: Erlangen, Germany, http://www.mol-net.com (accessed Sept. 9, 2013).

(31) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(32) Pastor, M. Alignment-Independent Descriptors from Molecular Interaction Fields. In *Molecular Interaction Fields. Applications in Drug Discovery and ADME predictions*; Cruciani, G., Ed.; Wiley−VCH: Chichester, U.K., 2006; pp 117−141.

(33) Mevik, B.-H.; Wehrens, R. The Pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Software* **2007**, *18*, 1−24.

(34) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.

(35) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.

(36) Guha, R. Chemical Informatics Functionality in R. *J. Stat. Software* **2007**, *18*, 1−16.

(37) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity—A Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.

(38) Maggiora, G. On Outliers and Activity Cliffs—Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(39) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(40) Wold, S, Johansson, E. C. M. PLS—Partial Least Squares Projections to Latent Structures. In *3D-QSAR in Drug Design, Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 523−550.