JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

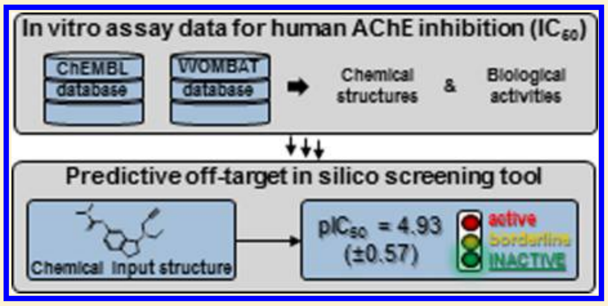
Generalized Workflow for Generating Highly Predictive in Silico Off-Target Activity Models

Lennart T. Anger,[†,‡] Antje Wolf,[†] Klaus-Juergen Schleifer,*[,†] Dieter Schrenk,[‡] and Sebastian G. Rohrer[§]

[†]Computational Chemistry and Biology, BASF SE, Carl-Bosch-Strasse 38, 67056 Ludwigshafen, Germany
[‡]Food Chemistry and Toxicology, University of Kaiserslautern, Erwin-Schroedinger-Strasse 52, 67663 Kaiserslautern, Germany
[§]Mechanistic Biology Fungicides, BASF SE, Speyerer Strasse 2, 67117 Limburgerhof, Germany

S *Supporting Information*

ABSTRACT: Chemical structure data and corresponding measured bioactivities of compounds are nowadays easily available from public and commercial databases. However, these databases contain heterogeneous data from different laboratories determined under different protocols and, in addition, sometimes even erroneous entries. In this study, we evaluated the use of data from bioactivity databases for the generation of high quality in silico models for off-target mediated toxicity as a decision support in early drug discovery and crop-protection research. We chose human acetylcholinesterase (hAChE) inhibition as an exemplary end point for our case study. A standardized and thorough quality management routine for input data consisting of more than 2,200 chemical entities from bioactivity databases was established. This procedure finally enables the development of predictive QSAR models based on heterogeneous in vitro data from multiple laboratories. An extended applicability domain approach was used, and regression results were refined by an error estimation routine. Subsequent classification augmented by special consideration of borderline candidates leads to high accuracies in external validation achieving correct predictive classification of 96%. The standardized process described herein is implemented as a (semi)automated workflow and thus easily transferable to other off-targets and assay readouts.



INTRODUCTION

Target proteins are large biomolecules such as receptors or enzymes relevant in a disease context or for the survival of a pathogen. In pharmaceutical or agrochemical research, small molecules are developed to modulate a specific target in order to achieve a desired effect. These selected targets are termed "on-targets". However, bioactive small molecules usually are not completely specific.[1] In pharmacology, for instance, adverse drug reactions are often caused by unwanted modulation of secondary targets. If a drug binds to another, unintended target, this target is a so-called "off-target".[1,2] Modulation of such an off-target might lead to potentially harmful (i.e., toxic) effects.

In early development of drug discovery[3] and crop-protection research, computer-based predictions and in vitro approaches are valuable methods for the screening of large numbers of ligand structures to reduce time and costs. Both, in vitro and in silico methods, are key instruments to achieve the goal of reducing animal experiments in the course of the "3R" concept (Replacement, Reduction, and Refinement).[4]

In pharmaceutical research, in vitro assays are routinely conducted to assess the inhibitory potential of new chemical entities for classical off-targets like hERG.[5] Semiautomated or fully automated high throughput in vitro assays are widely used screening tools. However, in silico methods usually offer an even higher throughput compared to in vitro assays, both in terms of the number of compounds and targets screened. In addition, even the potential of hypothetical molecules that have never been synthesized can be rapidly assessed—before investing resources in their synthesis. One typical field of application for computational methods is prioritization of compounds for further development or experimental testing in early research phases. For this purpose, structure—activity relationships for classic toxicological end points (e.g., mutagenicity, carcinogenicity)[6] or selected off-targets like PXR (pregnane X receptor)[7] or hERG[8,9] are determined.

For screening purposes, quantitative structure—activity relationship (QSAR) approaches are the most common in silico methods for toxicological end points. These techniques deal with the computer-assisted determination of relationships between chemical structures and their effects in biological systems. The benefit of QSAR models is the prediction of biological activities of new chemicals that have not been used in model development.[10] To establish predictive QSAR models, structure information on chemical compounds linked with bioactivity data is required. This data is used to train the QSAR model that describes the relationship between descriptors (derived from chemical structures) and biological activity. However, it must be stressed that the quality of input data is crucial for high-value prediction models. Within this context,

Received: June 10, 2014
Published: August 19, 2014

the source of the data used for model generation plays an essential role: Well-standardized assays like in-house assays or external assays from contract research organizations typically produce high quality data. Such homogeneous data, at best biological single-protocol data from congeneric series, show small variances and are therefore well suited for building models with high correlation.[11] On the other hand, the quantity of biological data coming from a single laboratory as well as the chemical space covered is often limited. Models derived from such homogeneous data often show poor predictive ability for more diverse new structures.[12]

Another valuable source for biological activity data has become increasingly popular in recent years: Chemical structures and associated bioactivities measured in various assays are nowadays easily available from large public databases like PubChem,[13] ChEMBL,[14] and DrugBank[15] or commercial sources like Liceptor[16] and WOMBAT.[17] However, studies have shown that a significant proportion of bioactivity database entries contain errors.[18,19] Therefore, a particular focus has to be placed on accurate data preparation and curation before model development, if data from bioactivity databases are used.

Biological data are subject to experimental and/or other types of error—this also applies for measurements conducted in one single laboratory. Variability is even greater if biological data is compiled from multiple laboratories.[11] In any case, it must be ensured that the compounds' mechanism of action is the same, so that models are finally built for a single defined end point having a mechanistic basis.[10] These known disadvantages are however outweighed by the extension of the chemical space covered by the model and the resulting increased external predictivity.[12]

Depending on the data sources used, there are basically two types of QSAR modeling strategies: So-called "explanatory QSAR models" are built with homogeneous data showing high correlation and statistical significance.[20] Such "local models" are developed to explain structure—activity relationships and possible trends within a data set of congeneric series. On the other hand, there are QSAR models developed for screening purposes covering a broad chemical space. Such "global" models are usually applied in virtual screening to discover new biologically active molecules.[20,21] However, these global models can also be used to assess whether compounds are likely to be active at certain "off-targets". In silico techniques could even provide information about structural properties associated with high off-target activity. At this early stage, chemists could make use of this information in order to redesign their compounds.

Each QSAR project consists of the following tasks: (i) accurate data preparation and curation, (ii) model development, and (iii) in-depth model validation. As mentioned above, quality of input data is crucial for high-value prediction models. Nevertheless, after model building and before taking QSAR models into practical operation one has to thoroughly evaluate the model's performance. There are two main concepts of model validation: Internal validation is typically used for model selection (i.e., tuning parameter optimization).[22] The model's generalizability or external predictive ability is assessed by external validation using data not taken into account during model development. Formerly, it was common practice to perform leave-one-out cross-validation as internal validation procedure. Meanwhile, it is widely accepted that this type of cross-validation leads to overoptimistic results for model predictivity.[23] Now it is generally recognized that external validation is the only proper way to establish a reliable QSAR model and to ensure good predictivity.[10,11,24] In order to answer the essential question how a QSAR model performs prospectively for new data, the overall data set is divided into two parts: The majority of the data is used as a modeling set for model development and can further be divided into training and internal test set(s). The small remaining portion of the entire data set is put aside as external validation set. This set is used at a later stage to characterize the model's behavior when faced with new data thus assessing the real "external" predictivity.[21]

During model validation with new data, one should bear in mind that every reference data set is (only) a sample from the chemical space. It is clear that a QSAR model can only capture structure—activity relationships supported by its training data. Thus, every model has its so-called "applicability domain" (AD) defined by the range of molecular descriptors used in model building and by the modeled response. Various approaches to express the scope and limitations of a model through definition of an AD were published by Netzeva et al.[25] An excellent overview about applicability domains especially for classification problems was given by Sushko et al.[26]

A special approach was recently published by Sheridan.[27] He proposed the application of a so-called "extrapolation curve", which is described as a graph showing the absolute prediction error as a function of the similarity of the compound under study to the nearest molecule in the training set.[27] Of course, every prediction is associated with a certain error. The method described by Sheridan permits to estimate the absolute prediction error (|predicted−observed|) for a given input structure. If the estimated error of prediction for a particular compound is above a defined threshold, the compound can be considered as out-of-domain because no reliable prediction can be made.

Within this study, human acetylcholinesterase (hAChE) inhibition serves as an example for classic (unwanted) off-target activity. AChE is a key-enzyme in the regulation of the levels of the neurotransmitter acetylcholine in the nervous system. As a serine hydrolase, AChE catalyzes the hydrolysis of acetylcholine and discontinues its synaptic transmission.[28] Exogenous compounds inhibiting AChE prevent the normal degradation of acetylcholine. This effect is specifically intended in the treatment of neurodegenerative diseases (e.g., Alzheimer's disease) where the hAChE serves as pharmaceutical "on-target" being reversibly inhibited.[28]

On the other hand, inhibition of acetylcholinesterase gained sad fame through organophosphate poisoning by chemical warfare nerve agents like sarin gas[29] or organophosphorus insecticides like parathion or paraoxon, respectively.[30] Here, the irreversible inhibition of the esterase leads to accumulation of acetylcholine in the synaptic cleft triggering continuous stimulation of nicotinic and muscarinic acetylcholine receptors. Symptoms of poisoning after acute exposure are muscle cramps, tachycardia, miosis, and respiratory paralysis. Also chronic exposure is known to cause severe neurotoxic effects.[28]

In vitro methods to assess acetylcholinesterase inhibitory activity of chemical compounds are well established. The most widely used technique is the Ellman Esterase Assay,[31] which allows tracking the formation of thiocholine from acetylthiocholine. Also several quantitative structure—activity relationship studies for hAChE inhibitory activity have been published.[32] However, previous work has focused on therapeutically used hAChE inhibitors with small data sets up to about 400 compounds.[33] To our knowledge, there is no large-scale QSAR study for hAChE inhibitory activity published yet. There remains a need for screening methods to detect hAChE inhibitory potential. Thus, hAChE inhibition was chosen as end point for our case study to evaluate the use of in vitro data from bioactivity databases for generation of high quality in silico models for off-target mediated toxicity.

One main outcome of our work is a standardized and thorough quality management routine for input data from bioactivity databases. In a straightforward modeling workflow, combinations of several descriptor sets and machine learning techniques for regression models were evaluated. The whole approach is completed through an extended applicability domain definition, which enables the development of predictive QSAR models based on public in vitro data from various sources. Finally, an in-depth validation using external validation sets was carried out. The obtained in silico models are well suited for screening purposes in drug discovery or crop-protection research in order to prioritize further development and reduce animal testing. Successful application is demonstrated for the detection of hAChE inhibitory potential as potential off-target activity. The standardized process described herein is implemented as a (semi)automated workflow and thus easily transferable to other off-targets and assay readouts.

■ METHODS

2.1. Data Preparation. For the work described in this study, data were used from two of the most prominent bioactivity databases ChEMBL ver. 15 (European Molecular Biology Laboratory)[14] and WOMBAT (World of Molecular Bioactivity) ver. 2012.1 (Sunset Molecular).[17,34] However, prior to the actual modeling task, the data provided by bioactivity databases have to be prepared. Therefore, special emphasis was put on accurate data preparation and curation (see Figure 1).

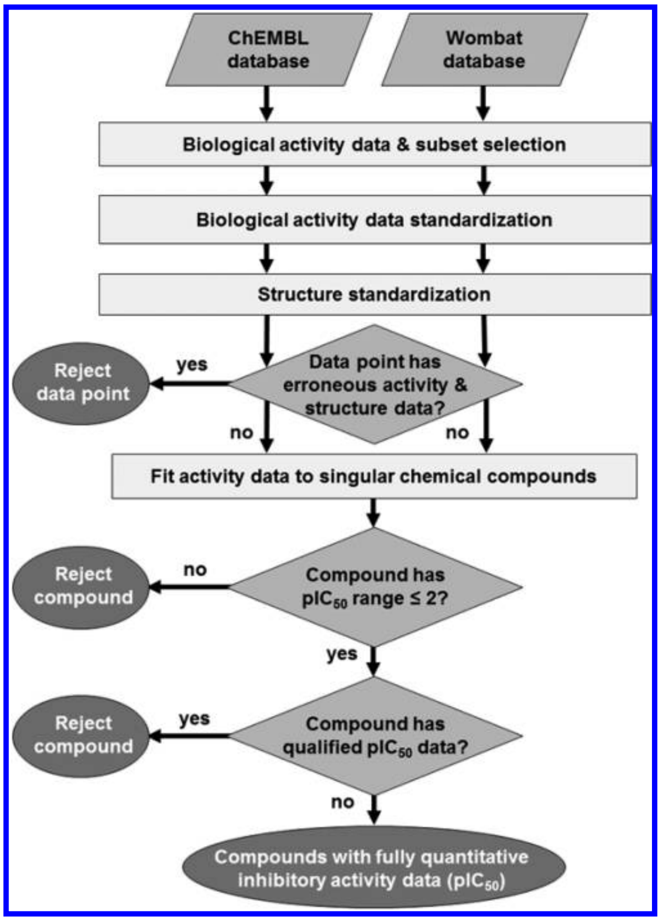

ChEMBL database
Wombat database
Biological activity data & subset selection
Biological activity data standardization
Structure standardization
Reject data point
yes
Data point has erroneous activity & structure data?
no
no
Fit activity data to singular chemical compounds
Reject compound
no
Compound has pIC_{50} range ≤ 2?
yes
Reject compound
yes
Compound has qualified pIC_{50} data?
no
Compounds with fully quantitative inhibitory activity data (pIC_{50})

Figure 1. Data set preparation workflow: 1) Biological activity data and subset selection [select target "human AChE", assay readout "inhibition, IC_{50}"], 2) Biological activity data standardization [IC_{50}-to-pIC_{50} conversion if necessary, flag qualified activity data], 3) Structure standardization [convert provided smiles to standardized 2D SD structures using OpenBabel, derive universal smiles with OpenBabel], 4) Check for erroneous activity and structure data, 5) Fit activity data to singular chemical compounds [group data points by universal smiles, aggregate individual data point information, keep median of all individual pIC_{50} values per compound], 6) Check compounds' merged pIC_{50} range, 7) Check compounds for qualified pIC_{50} data.

In both databases, ChEMBL and WOMBAT, "human acetylcholinesterase" was selected as target, and the selected readout was half-maximal inhibitory concentration (IC_{50}). Each extracted data point consisted of a chemical structure and an associated bioactivity for hAChE inhibition. Where necessary, IC_{50} values provided in various units were converted to pIC_{50} scale ($pIC_{50} = -\log_{10} IC_{50}$).

Ligand structure information was provided by both databases in SMILES[35] (Simplified Molecular Input Line Entry Specification) format. However, different software packages apply their own nonuniform canonicalization algorithm when generating canonical SMILES.[36] To ensure a reliable detection of replicate structures, the given SMILES strings were converted into a 2D structure-data (SD) format using OpenBabel 2.3.2.[36,37] Counterions were removed (all but the largest contiguous fragment is eliminated), and hydrogens were added. These standardized 2D structures were subsequently converted with OpenBabel into "Universal SMILES", a method, which uses canonical labels from IUPAC's InChI[38] to build highly standardized canonical SMILES strings.[39]

At this stage, data points with structure data being not suitable for modeling purposes (i.e., inorganic metal salts) were filtered out. Moreover, some data points were manually blacklisted due to inconsistencies or errors in the extracted data (e.g., in section 3.1). For the remaining data points, activity data for the same chemical compound were pooled through grouping by Universal SMILES. The median of all individual grouped pIC_{50} values was used, and additional data point information was aggregated.

Compounds with pooled activity data within a pIC_{50} range of >2 (highest pIC_{50} − lowest pIC_{50} > 2) from the underlying original data points were considered unreliable and therefore excluded from further processing. Finally, for compatibility with all applied machine learning techniques, compounds with qualified activity data (e.g., "pIC_{50} < 4") were rejected.

2.2. Model Fitting. In order to obtain high-value QSAR models and to ensure their predictivity, a straightforward modeling workflow (see Figure 2) based on a two-layered validation approach (for simultaneous model selection and estimation of the generalization error) called "Two Deep" validation[40] was developed using KNIME[41] (version 2.7.2) and R Environment for Statistical Computing[42] (version 2.10.1).

Data Set Division. Data set division has been an intensely discussed topic for the last years.[23] The most frequently used method is the random division of the entire data set. A contrary approach is based on the rational splitting of the data with the aim of diversity selection (using sphere exclusion algorithms for example).[43] Here, random splitting was performed based on advantages of this approach as discussed in a recent study.[23]

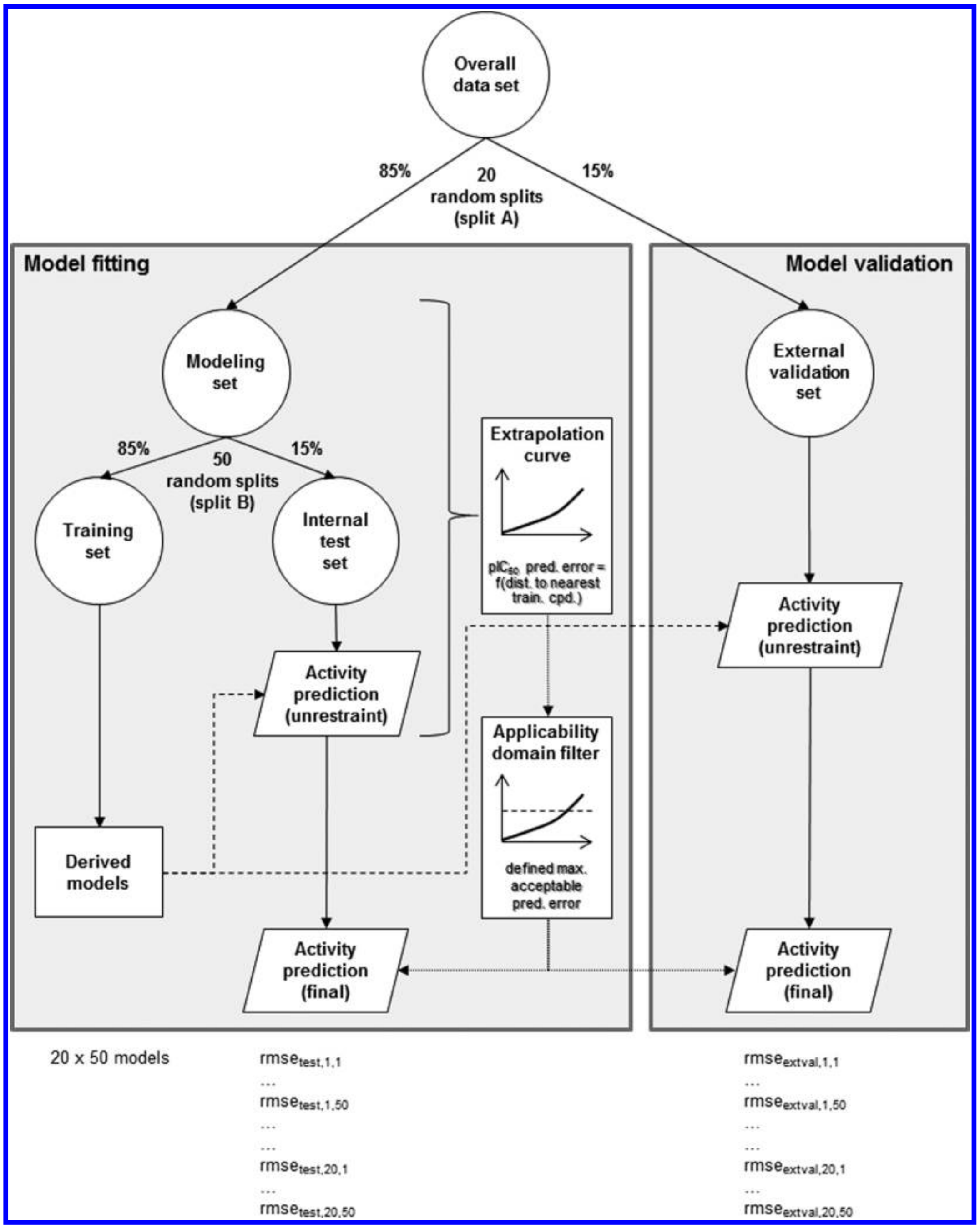


Figure 2. Modeling workflow ("Two Deep" validation): First the entire data set is split into two portions of different sizes: For each split A, the majority of the data (85%) is used for "model fitting" (left part), whereas the remaining 15% are kept for "model validation" (right). Within "model fitting", the remaining data is again randomly split 50 times (split B) into training (85%) and internal test (15%) sets. Prediction results of internal test set molecules are taken into account in construction of the extrapolation curve, which is used as error estimation routine and for the definition of the applicability domain.

At first, the entire data set is split into two portions of different sizes: For each split A, the majority of the data (85%) is used for "model fitting" (left part of the workflow diagram, Figure 2), whereas the remaining 15% are kept for "model validation" (right section, Figure 2). This random splitting is repeated 20 times in an outer loop to reduce bias. Data splitting was performed in KNIME with different random seeds.

The "model fitting" part contains an additional inner loop, where the remaining data is randomly split 50 times (split B) into training (85%) and internal test (15%) sets. Inside this inner loop, best parameter settings are identified using (average) root-mean-square error of internal test set predictions as quality measure. Since prediction results of internal test set molecules are taken into account in construction of the extrapolation curve, structures are involved in the model fitting process and cannot be considered as "new" compounds for a proper validation.

Descriptors. For this study, the two most widely used 2D descriptor packages were chosen: Dragon (version 6.0, Talete srl, Milan, Italy)[44] and Molecular Operating Environment (MOE, version 2012.10, Chemical Computing Group Inc., Montreal, Canada).[45] A set of 4252 Dragon descriptors was initially calculated for the overall prepared data set. This set contains all standard 2D descriptors implemented in Dragon version 6.0. It was further reduced to 2246 descriptors by removing constant descriptor columns as well as descriptors with missing values. Second, MOE was used to calculate 187 standard 2D descriptors, whereby partial atomic charges were adjusted internally using MMFF94*.[46]

For each training set, the descriptor matrices of the two descriptor sets were standardized by z-transformation (columns were normalized to a mean value of zero and a standard deviation of one) in KNIME. Within this subset, constant descriptor columns having a standard deviation of zero were removed. Accordingly, the same descriptor columns were removed from the descriptor matrices of the corresponding internal test and external validation sets. The scaling of the remaining descriptor columns was transferred to the corresponding internal test and external validation sets. The number of descriptors available for model fitting varied slightly over different training sets, ranging from 2,243 to 2,246 for Dragon descriptors and from 185 to 187 for MOE descriptors.

In addition to the use of entire descriptor matrices as described above, Principal Component Analysis (PCA)[47,48] was applied as a preprocessing tool for data dimension reduction with both descriptor sets. Principal components (PCs) are linear combinations of the original variables and finally serve as new input for QSAR modeling. Beforehand, descriptor matrices were standardized (scaling, etc.) as described above. The PCA method was applied on every single training set descriptor matrix using "R". As many PCs were extracted as were needed to achieve a cumulative variance explained of 99%. Finally, PCA was applied analogously to each corresponding internal test set and external validation set.

Machine Learning Techniques. As it is generally assumed for QSAR tasks that there is no specific model class that always fits the data best,[20] several machine learning techniques for the regression task were evaluated. Partial Least Squares (PLS),[49] Random Forest (RF),[50] and Support Vector Regression (SVR)[51] were run under "R". As part of our case study, a manageable grid of tuning parameters was evaluated for each model class.

PLS regression is probably the most popular linear method in the field and widely used.[20,52] The PLS method was developed by Herman and Svante Wold[49] and is based on the projection to latent structures: principal component-like vectors are extracted from the descriptor data (Xi) as well as from the response data (Y). Finally, a linear regression equation is built from these factors. In model building, the PLS tuning parameter is the number of selected latent variables.[52] For our studies, kernel PLS (from R package "pls", version 2.3) was chosen together with its tuning parameters ncomp ∈ {1, 2, 4, 7, 9, 11, 13, 16, 18}. Compared to SVR and RF, PLS is less complex and needs considerably less computation time but cannot handle qualified activity data. For consistency reasons, compounds with qualified activity data were generally excluded from the final data set (see 2.1).

The Random Forest method was proposed by Leo Breiman in 2001.[50] It consists of an ensemble of decision trees. Recently, the RF algorithm was applied in QSAR tasks with increasing frequency.[53] Especially in the context of predictive toxicology[54,55] the usage of RF seems promising: RF is designed to handle high-dimensional data, even if containing high-order interactions or correlated predictor variables. The method does not require a variable preselection because trees containing irrelevant descriptors are wiped out. The method is robust with respect to the structure of the training data, for example multiple clusters and skewed distributions. It was even proposed that the RF algorithm is able to handle multiple mechanisms of action.[53] Within our workflow, we used the R package "randomForest" (version 4.6-7) and evaluated two tuning parameters for RF regression: the number of regression trees in the forest ("ntree") and the number of variables (features) that are randomly selected for each split during tree induction ("mtry"). A larger number of trees usually leads to a more stable and robust forest. For evaluation of ntree, we chose ntree ∈ {100, 500, 1,000}. The default for mtry in regression tasks is $p/3$, where p is the number of variables. Thus, we chose the mtry parameter according to the size of the training descriptor matrix. For our studies, mtry parameters were chosen as follows: for Dragon descriptors mtry ∈ {700, 1,000, 1,400}, for Dragon descriptors after PCA mtry ∈ {50, 80, 100}, for MOE descriptors mtry ∈ {60, 90, 120}, and for MOE descriptors after PCA mtry ∈ {8, 12, 16}.

Support Vector Machines were originally developed by Cortes and Vapnik for classification problems in the field of pattern recognition.[51] A Support Vector Machine tries to separate a set of data points according to their class labels by insertion of a hyperplane into the multidimensional descriptor space so that class boundaries are as widely separated as possible (maximum margin). The method is very flexible and can handle classification as well as regression tasks. In the latter case, one speaks of Support Vector Regression (SVR).[56,57] In contrast to PLS and RF, SVM/SVR models are of high complexity and cannot be interpreted.

Within this study, we used the R package "e1071" (version 1.6-1) for Support Vector Regression. For the radial base kernel function ("RBF kernel") we evaluated the tuning parameters cost (c), epsilon (ε), and gamma (γ): c is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error [we chose $\log_{10}(c) \in \{0, 1, 2\}$]; ε is a parameter of the ε-insensitive loss function that affects the number of support vectors used to construct the regression function [$\log_{10}(\varepsilon) \in \{-2, -1\}$]; and γ stands for the width of the radial basis function. For evaluation of γ, parameters were chosen as follows: for Dragon descriptors $\log_{10}(\gamma) \in \{-4, -3\}$, for Dragon descriptors after PCA $\log_{10}(\gamma) \in \{-3, -2\}$, for MOE descriptors $\log_{10}(\gamma) \in \{-3, -2\}$, and for MOE descriptors after PCA $\log_{10}(\gamma) \in \{-2, -1\}$.

Error Estimation Routine. Regression models were used to predict the pIC_{50} activities of the corresponding internal test set molecules for each parameter combination of each model class. Predicted and observed activity values of the internal test set compounds are used to establish an "error estimation routine". The fundamental basis of this approach is the extrapolation curve proposed by Sheridan.[27] This method permits the estimation of the absolute prediction error (|predicted− observed|) of a given input structure. Whereas the original implementation operates with overlapping bins, we used a smoothing function within our modified version of the extrapolation curve.

Pairwise dissimilarities in the descriptor space were calculated as Euclidean distances using "R". On the basis of the resulting distance matrices, the distance to the nearest molecule in the corresponding training set was determined for each individual internal test set molecule. The absolute prediction errors (|pIC_{50}[predicted] − pIC_{50}[reported]|) of all internal test set molecules and the associated distances to their nearest training

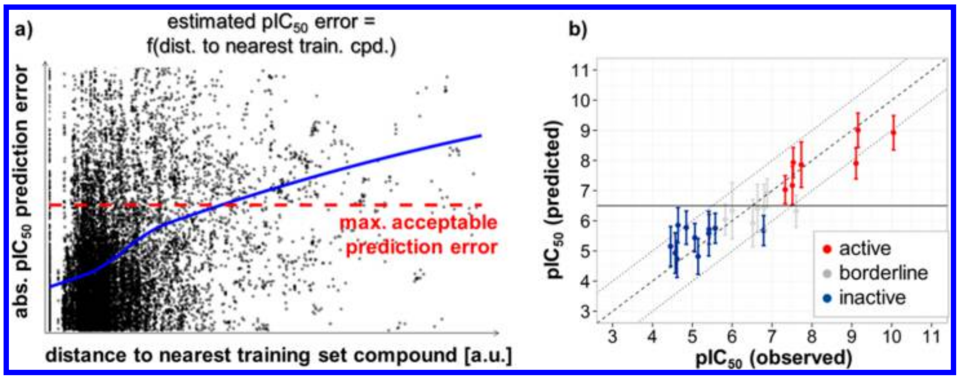


Figure 3. a) Exemplary depiction of an extrapolation curve from the best SVR model: Predictions of internal test set compounds are used in the development of extrapolation curve (inspired by Sheridan[27]). Method allows estimating absolute prediction error (|predicted−observed|) of given input structure. b) Simplified exemplary scatterplot with error bars symbolizing the uncertainty of predictions (± estimated absolute prediction error): Compounds considered out-of-domain, if estimated error of prediction is above the threshold of 1 on the pIC_{50} scale. Compounds are labeled "borderline" if activity class allocation uncertain due to their estimated pIC_{50} errors (error bars crossing the threshold line).

molecule were combined for all split B's (within one split A) and were jointly fitted into a LOESS smoothing function (see depiction in Figure 3a) in "R". Finally, one model for each split A and parameter set was obtained providing an estimate of the "typical" absolute prediction error[27] that depends on the distance of the compound under study to the nearest molecule in the training data.

If an internal test set compound is dissimilar from compounds of the training data (in terms of descriptor data) the model has to extrapolate beyond the underlying data range. Predictions for compounds outside this applicability domain are not necessarily invalid but at least less reliable. In practice, the detection of test compounds, for which the model cannot generate reliable predictions, is crucial. Instead of setting a fixed empirical threshold value for the maximum acceptable distance to a training set compound in the descriptor space, we use the error estimation routine described above to obtain details about the extent of extrapolation needed as well as the resulting expected prediction error. We defined a threshold of 1 as (absolute) maximum acceptable estimated prediction error on the pIC_{50} scale (see Figure 3a). Thus, if the estimated error of prediction for a particular compound is above this threshold (higher distance to nearest training set compound than distance corresponding to an expected pIC_{50} error of 1), this molecule is considered as out-of-domain. Because no reliable prediction of its bioactivity can be made, the molecule is excluded from further processing.

Classification. In the first place, a reliable assessment is needed in early routine screening whether compounds are likely to be active at a certain "off-target" like the hAChE. Therefore, in addition to the fully quantitative activity prediction, a subsequent classification of the numeric predictions was carried out. A pIC_{50} of 6.5 was chosen as threshold for classification being the rounded median of the activity distribution of this data set. Compounds having a $pIC_{50} \leq 6.5$ would be referred to as "inactives", and, accordingly, compounds with a $pIC_{50} > 6.5$ are considered as "actives".

However, a sharp distinction between actives and inactives is not always possible: Like in vitro experiments, also in silico predictions, derived from experimental data, are associated with a certain error. Within this study, the error of each individual prediction on the pIC_{50} scale was estimated using the modified extrapolation curve. Consequently, for classification purposes, it has to be checked whether the estimated pIC_{50} error of each compound can affect its attribution to one of the two activity classes. Compounds were labeled as "borderline", if the activity class allocation is uncertain due to their estimated pIC_{50} errors (see Figure 3b).

2.3. Model Validation.

In the "Model Validation" section of the workflow, the essential question was evaluated how the created QSAR models (developed in the "model fitting" part) perform prospectively for new data. Repeating the random split A for 20 times, 20 different modeling/external validation set partitions were generated. For each external validation set, activities were predicted individually using the corresponding 50 QSAR models derived from training sets originating from 50 split B's. After initial unrestrained predictions, the described AD approach was applied, and prediction errors were estimated using the extrapolation curve. Quality measures are given as average of all 1,000 data set splits for internal test sets and external validation sets.

For quantitative predictions, the coefficient of determination (R^2) was used as measure of goodness-of-fit. The root-mean-square error (RMSE) as well as the mean absolute error of prediction (MAE) indicate the deviations of the predicted from the experimental values.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_{i,obs} - y_{i,pred})^2}{\sum_{i=1}^{N} (y_{i,obs} - \overline{y_{obs}})^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_{i,obs} - y_{i,pred})^2}$$

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_{i,obs} - y_{i,pred} \right|$$

Here, $y_{i,obs}$ is the observed and $y_{i,pred}$ is the predicted activity of compound i, whereas $\overline{y_{obs}}$ is the average observed activity within the training and validation sets, respectively. In the particular case of RF, $y_{i,pred}$ represent averages from ensembles of regression trees. Additionally, the performance of subsequent classification was evaluated using confusion matrices and according "Cooper statistics".[58] The sensitivity (SN) expresses the ability to correctly predict positive/active compounds as active:

$$\mathrm{SN} = \frac{\text{no. of true positives}}{(\text{no. of true positives} + \text{no. of false negatives})}$$

The specificity (SP) indicates the ratio of inactive compounds correctly predicted as inactive:

$$\mathrm{SP} = \frac{\text{no. of true negatives}}{(\text{no. of true negatives} + \text{no. of false positives})}$$

As general measure for the correctness of predictions, the measure of "correct classification rate" alias CCR (sometimes also called "balanced accuracy") is used:

$$\mathrm{CCR} = \frac{(\mathrm{SN} + \mathrm{SP})}{2}$$

■ RESULTS AND DISCUSSION

3.1. Data Set Information. The established data preparation workflow was used to process chemical structures and associated human AChE inhibition values from the two most prominent bioactivity databases, ChEMBL and WOMBAT. At the first filtering step of the workflow, data points associated with inorganic metal salts like CHEMBL2097081 (sodium nitroprusside) were removed due to suspicion on rather nonspecific interactions in the biological test systems. Moreover, some data points (see the full list in the Supporting Information, Table S4) were blacklisted due to inconsistencies or errors in extracted data. For example, within the set extracted from ChEMBL_15, the data point corresponding to compound ID "CHEMBL636" had to be rejected because it referred to a secondary end point ("Inhibition of AChE-induced amyloid beta aggregation").

When activity data for the same chemical compound were pooled, the median of all individual pIC_{50} values for this compound was finally preserved. We chose the median as a robust measure of central tendency here. In a similar case, it has been proposed to take the arithmetic average of similar experimental properties.[59]

Primary data for compounds having pooled activity data with pIC_{50} range >2 were manually inspected. The range filter criterion was of particular value here: An observational error of about ± one log unit is not uncommon in most IC_{50} biological assay data,[19] and thus compounds having a pIC_{50} range ≤ 2 were considered acceptable in the context of this study. However, compounds with a pIC_{50} range with a value of about 3 most likely indicated unit conversion errors during the data mining process for at least one of the underlying data points.[19,60] To give an example, in the ChEMBL_15 compound CHEMBL191461 was specified with an IC_{50} of 1,800 nM (pIC_{50} = 5.74), whereas the same molecule was listed in WOMBAT (SMDL-00153084) with a pIC_{50} of 8.74. Manual inspection of the common primary literature[61] revealed that the original value of 18 nM was transferred incorrectly into ChEMBL_15. Consequentially, all data points in ChEMBL originating from this publication were blacklisted using the corresponding assay ID (CHEMBL832454).

After thorough data preparation and curation, 2,203 compounds reported with fully quantitative hAChE inhibitory activity were finally obtained and formed the data basis for this study. Comprised molecules are chemically diverse with molecular weights between 61 and 898 Da and have 0 to 33 rotatable bonds and a logP(o/w) ranging from −0.82 to 12.68. On the pIC_{50} scale, the activity ranged from 2.9 to 10.7. Thus, nearly 8 orders of magnitude in IC_{50} activity are covered. A corresponding activity distribution histogram is shown in Figure 4. Within the final hAChE data set, about 97% of all compounds with merged activity data (478 molecules) have an IC_{50} range smaller than one log unit. Analysis of the database sources of all molecules revealed that bioactivity data for 1,495 compounds (68% of overall set) descended from the ChEMBL database only (see Venn diagram in Figure 5). Data for 416 molecules (19%) were extracted from both bioactivity databases. This means that the overlap of the commercial database WOMBAT and the public ChEMBL was rather large, because 416 out of 708 compounds extracted from WOMBAT were already covered by ChEMBL. However, additional 292 molecules with associated bioactivities (13%) could be uniquely obtained from WOMBAT.

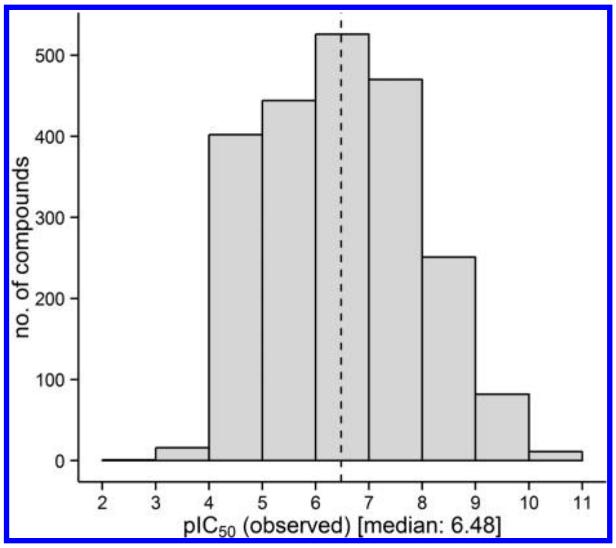


Figure 4. Activity distribution histogram of pIC_{50} values from 2203 compounds of the overall prepared hAChE data set: Values are ranging from 2.9 to 10.7, thus covering nearly 8 orders of magnitude in IC_{50} activity.

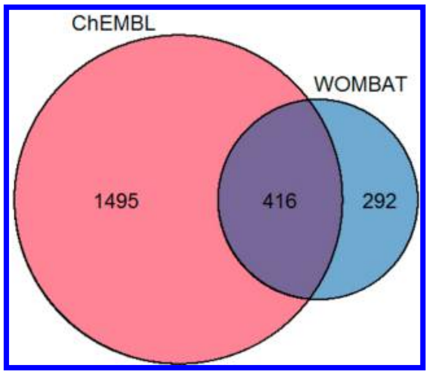


Figure 5. Venn diagram showing counts of molecules from the overall prepared hAChE data set according to their data origins: bioactivity data for 1495 compounds are descended from the ChEMBL database only, 292 molecules with associated bioactivities could be uniquely obtained from WOMBAT. Additionally, there is an overlap of 416 molecules for which extracted data comes from both bioactivity databases.

3.2. Prediction Results for Regression Task. Within this case study, 360,000 models for collections of 1,000 training/test sets with two descriptor sets, two different ways of descriptor preprocessing, three regression algorithms, and 30 parameter settings (12 for SVR, 9 for RF, and 9 for PLS models) have been developed in total. All quality measures shown are average results for 1,000 data set splits for internal test sets comprising of 281 molecules and external validation sets containing 330 compounds.

Regression results for each model class and descriptor set combination are listed in Table 1. Reported are quality measures for parameter settings that showed smallest average root-mean-square error for internal test set predictions. Results for all tested tuning parameter combinations can be found in the Supporting Information (Table S5).

Table 1. Regression Results - Averaged Quality Measures for 1000 Data Set Splits: The Coefficient of Determination (R^2), Root-Mean-Square Error (RMSE, pIC_{50} scale), and Applicability Domain Coverage[a]

method	internal test sets			external validation sets		
model/ descriptors	R^2	RMSE	coverage [%]	R^2	RMSE	coverage [%]
SVR/D	0.713	0.749	95	0.712	0.751	95
SVR/D (P)	0.699	0.768	97	0.698	0.769	96
SVR/M	0.694	0.772	98	0.694	0.773	98
SVR/M (P)	0.664	0.809	98	0.664	0.811	97
RF/D	0.695	0.773	99	0.688	0.782	99
RF/D (P)	0.632	0.850	96	0.627	0.855	96
RF/M	0.691	0.778	98	0.684	0.787	98
RF/M (P)	0.619	0.864	97	0.615	0.868	97
PLS/D	0.568	0.918	91	0.561	0.924	91
PLS/D (P)	0.494	0.997	95	0.494	0.995	95
PLS/M	0.466	1.020	93	0.467	1.018	93
PLS/M (P)	0.311	1.141	68	0.312	1.147	69

[a] SVR: Support Vector Regression, RF: Random Forest, PLS: Partial Least Squares, D: Dragon, M: MOE, (P): Principal Component preprocessing.

For the regression task, the best overall results were obtained using Support Vector Regression with the full set of Dragon descriptors. On average, these models showed the highest coefficient of determination (R^2) accounting for 71.3% of variance in observed activities in the internal test sets and 71.2% in external validation. Furthermore, they showed the overall lowest average root-mean-square error (RMSE) of 0.749 on the pIC_{50} scale for the test sets and 0.751 for external validation sets with a domain coverage of 95% in both cases (Table 1). Thus, quality measures for internal test sets can be considered predictive and meaningful for external validation results. SVR models built with MOE descriptors showed slightly lower R^2 and somewhat higher RMSE values compared to Dragon descriptor-based models. PCA as a preprocessing tool used for data dimension reduction caused a decreased predictive performance compared to both full descriptor set approaches. Comparison of all SVR results revealed that quality measures strongly depend on the tuning parameters chosen for these models. For different tuning parameters of the predefined grid, coefficients of determination for test set predictions using Dragon descriptors, for example, ranged from 0.486 to 0.713. The best performance was achieved using $c = 10$, $\log_{10}(\gamma) = -3$, and $\varepsilon = 0.1$ for both descriptor sets. Their preprocessed sets having less dimensions performed better with a reduced γ value [$\log_{10}(\gamma) = -4$].

Using the Random Forest algorithm, however, quality measures for different parameter combinations evaluated for the individual descriptor sets varied very little. This demonstrates the robustness of Random Forests.[53] Random Forest models based on the full Dragon descriptor set showed the highest overall coverage (99%). The R^2 was 0.695 with an RMSE of 0.773 for test sets and an R^2 of 0.688 with RMSE of 0.782 for the external validation sets. MOE descriptor-based models were somewhat less predictive compared to Dragon descriptor-based models. Preceding data dimension reduction by PCA for both descriptor sets reduced predictive performance. One of the Random Forest algorithm's specific strengths is its own implicit feature selection. This is why RF models might perform so well with virtually unprepared and full Dragon and MOE descriptor sets.

PLS models were in general significantly less predictive in our case study. The highest R^2 for internal test sets was achieved with MOE descriptors (0.568). These models accounted only for 56.1% of variance in observed activities in external validation sets. This is hardly surprising, because it is well-known that PLS cannot handle noisy data very well. Besides, it is limited to less complex interactions and single mechanisms. Although the PLS algorithm already does an implicit dimension reduction, descriptor sets preprocessed by PCA were also used as input for PLS modeling for consistency reasons. It can be seen that this procedure leads to a tremendous loss of information. Especially for MOE descriptor sets the coverage was very low (68% and 69%) and the R^2 values were generally poor.

Many AD approaches focus on coverage of training set or descriptor space only. In contrast, our approach takes into account the model itself and its uncertainty. We established an error estimation routine based on Sheridan's extrapolation curve[27] to obtain details about the extent of extrapolation needed and the resulting expected prediction error for each queried compound. A threshold of 1 as maximum acceptable estimated prediction error on the pIC_{50} scale was defined as upper limit for the extent of extrapolation. Thus, if internal test or external validation compounds have a higher distance to their nearest training set molecules than the distance corresponding to an expected pIC_{50} error of 1, these compounds are considered as out-of-domain. Applicability domain coverage for the best prediction models shown in Table 1 generally was between 91 and 99%, except from the aforementioned PLS model with preprocessed MOE descriptor sets (69%).

Reported values in Table 2 show nicely that, on average, the estimated prediction errors coming from the extrapolation curve (mean estimated error of prediction, MEE) are in good agreement with MAE (mean absolute error of prediction). The error estimation routines were developed on the basis of internal test set predictions. However, even for external validation sets the estimated prediction errors aligned well with the true absolute error of predictions.

SVR models based on Dragon descriptors showed smallest MAE for internal test sets (0.536) and external validation sets (0.534). The mean estimated error of prediction in these cases was 0.533 or 0.570, respectively.

3.3. Prediction Results for Classification Task. Besides the basic quantitative activity prediction, a subsequent classification based on the regression results was carried out. Within this case study, a pIC_{50} of 6.5 was chosen as classification cutoff being the rounded median of the activity

Table 2. Regression Results - Averaged Quality Measures for 1000 Data Set Splits: Mean Absolute Error of Prediction (MAE, pIC_{50}) and Mean Estimated Error of Prediction (MEE, pIC_{50}) Obtained from the Established Error Estimation Routine for Compounds Being in Domain[a]

method	internal test sets		external validation sets	
model/descriptors	MAE	MEE	MAE	MEE
SVR/D	0.536	0.533	0.534	0.570
SVR/D (P)	0.549	0.547	0.548	0.544
SVR/M	0.556	0.558	0.559	0.557
SVR/M (P)	0.594	0.594	0.598	0.593
RF/D	0.577	0.583	0.582	0.581
RF/D (P)	0.657	0.655	0.658	0.654
RF/M	0.579	0.583	0.585	0.582
RF/M (P)	0.668	0.668	0.670	0.669
PLS/D	0.710	0.706	0.710	0.705
PLS/D (P)	0.789	0.783	0.783	0.782
PLS/M	0.809	0.808	0.809	0.808
PLS/M (P)	0.928	0.926	0.933	0.926

[a] SVR: Support Vector Regression, RF: Random Forest, PLS: Partial Least Squares, D: Dragon, M: MOE, (P): Principal Component preprocessing.

distribution of the overall data set (see Figure 4). Consequently, data subsets obtained after random splitting are fairly balanced with respect to the number of actives and inactives. In practical applications, however, one could use a threshold value of a known effect as classification threshold instead - or the activity of a reference compound.

The results of the subsequent classification, corresponding to regression models introduced in the last section (see Table 1), are summarized in Table 3: On average, the best prediction method obtained from SVR models and Dragon descriptors correctly predicted 85.1% of the AChE inhibitors and 85.4% of the noninhibitors in the internal test sets. In the external validation sets 85.3% of the actives and 85.5% of the inactives were predicted correctly. Like already shown for the regression task, applicability domain coverage was 95% in both cases. A correct classification rate of 85.3% is considerable for such a huge and inhomogeneous data set. Approaches based on SVR models with preprocessed Dragon descriptor sets, MOE descriptor sets, or preprocessed MOE descriptor sets were slightly less predictive in classification. Classification methods based on developed Random Forest regression models showed an increased specificity at a higher coverage compared to their SVR analogues built with the same descriptor sets. The best RF-based classification approach was assembled with Dragon descriptors showing a sensitivity of 83.8% and a specificity of 86.1% in internal test sets or 83.4% (SN) and 86.0% (SP) in external validation. PLS predictions produced best results with MOE descriptor sets: 80.7% sensitivity and 77.5% specificity.

In addition to the standard classification approach, we introduce here our borderline classification approach which takes into account the methods' uncertainty around the classification threshold. Estimated pIC_{50} errors derived from the extrapolation curve were used for detection and labeling of "borderline compounds". If activity class allocation of compounds was uncertain due to their estimated pIC_{50} errors (compounds' predicted pIC_{50} ± estimated absolute prediction error crossing the threshold line of 6.5 on the pIC_{50} scale, see Figure 3b), these compounds were labeled as "borderline" and excluded from classification (see Figure 6).

Using this borderline classification approach, the classification performance significantly improved: For the aforementioned best-performing combination of SVR models and Dragon descriptors coverage dropped from 95% to 64% due to exclusion of borderline candidates. However, these models then correctly predicted 94.6% of the AChE inhibitors and 92.4% of the noninhibitors in the test sets within its applicability domain. 93.6% of the actives and 94.8% of the inactives were predicted correctly in external validation. Prediction methods based on Random Forest regression models also showed even higher correct classification rates at the expense of decreased coverage. For example, RF models developed on preprocessed Dragon descriptors reached 96.0% correct classification rate at a coverage of 45.0% in internal test sets and 95.9% CCR at a coverage of 44.0% in external validation. Borderline classification approaches based on PLS regression models, however, did not appear meaningful. They showed much too small coverage (between 13 and 51%): First, because up to 32% of the compounds were considered out-of-domain having an estimated pIC_{50} prediction error >1 and, second, because of inaccurate quantitative predictions (see high

Table 3. "Borderline Classification Approach" Results - Averaged Quality Measures for 1000 Data Set Splits: Sensitivity (SN), Specificity (SP), Correct Classification Rate (CCR), and Applicability Domain Coverage[a]

method	internal test sets				external validation sets			
model/descriptors	SN [%]	SP [%]	CCR [%]	coverage [%]	SN [%]	SP [%]	CCR [%]	coverage [%]
SVR/D	94.6 (85.1)	92.4 (85.4)	93.5 (85.3)	64.0 (95.0)	95.2 (85.3)	92.4 (85.5)	93.8 (85.4)	64.0 (95.0)
SVR/D (P)	94.2 (85.3)	91.5 (84.3)	92.9 (84.8)	66.0 (97.0)	94.5 (85.0)	91.4 (84.2)	93.0 (84.6)	66.0 (96.0)
SVR/M	93.7 (84.2)	91.7 (83.9)	92.7 (84.0)	66.0 (98.0)	92.9 (83.3)	92.1 (84.2)	92.5 (83.7)	66.0 (98.0)
SVR/M (P)	93.4 (83.5)	90.3 (82.5)	91.8 (83.0)	64.0 (98.0)	93.0 (82.4)	90.8 (82.9)	91.9 (82.7)	64.0 (97.0)
RF/D	93.7 (83.8)	95.3 (86.1)	94.5 (84.9)	60.0 (99.0)	93.5 (83.4)	94.7 (86.0)	94.1 (84.7)	60.0 (99.0)
RF/D (P)	95.6 (81.6)	96.3 (85.3)	96.0 (83.4)	45.0 (96.0)	95.8 (81.9)	96.1 (85.6)	95.9 (83.8)	44.0 (96.0)
RF/M	93.7 (84.3)	95.0 (85.0)	94.3 (84.7)	59.0 (98.0)	93.4 (83.9)	94.7 (84.9)	94.1 (84.4)	58.0 (98.0)
RF/M (P)	94.0 (80.9)	94.6 (82.6)	94.3 (81.7)	46.0 (97.0)	94.1 (80.8)	94.8 (82.6)	94.4 (81.7)	45.0 (97.0)
PLS/M	91.5 (80.7)	90.7 (77.5)	91.1 (79.1)	51.0 (91.0)	91.1 (80.3)	91.7 (77.7)	91.4 (79.0)	50.0 (91.0)
PLS/M (P)	93.0 (78.5)	91.5 (76.6)	92.2 (77.5)	39.0 (95.0)	92.7 (78.5)	92.7 (76.2)	92.7 (77.4)	39.0 (95.0)
PLS/D	91.4 (75.4)	92.5 (75.1)	91.9 (75.2)	36.0 (93.0)	91.7 (74.8)	93.5 (75.7)	92.6 (75.2)	36.0 (93.0)
PLS/D (P)	94.6 (66.4)	91.7 (67.3)	93.2 (66.8)	13.0 (68.0)	94.3 (67.8)	93.3 (67.4)	93.8 (67.6)	13.0 (69.0)

[a] Values in parentheses indicate the quality measures when borderline compounds are also taken into account. SVR: Support Vector Regression, RF: Random Forest, PLS: Partial Least Squares, D: Dragon, M: MOE, (P): Principal Component preprocessing.

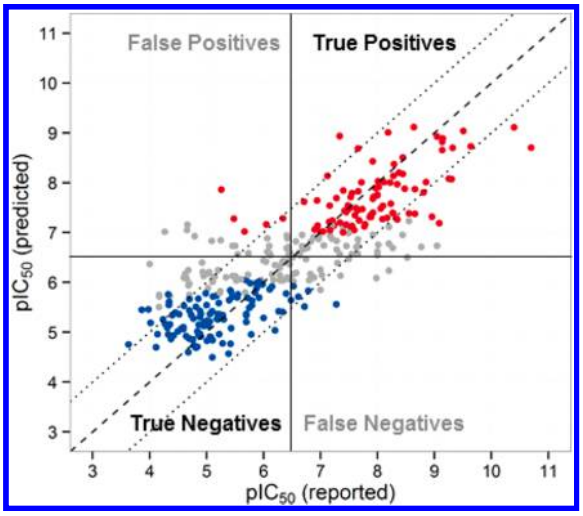


Figure 6. Exemplary scatterplot of one data set split illustrating the "borderline classification approach" based on the regression results applying a pIC_{50} threshold of 6.5.

RMSE in Table 1) causing excessive proportion of compounds with high estimated prediction errors around the classification threshold.

This borderline classification approach based on a rather unusual combination of regression and subsequent classification proved appropriate for practical applications for several reasons: The method provides a reliable and robust assessment of the compounds' off-target liability in the form of class label "active", "inactive", or "borderline". Assignment of "borderline" label is made based on the estimated pIC_{50} error of the underlying regression model. Apart from a class label, there is a fully quantitative prediction result together with its estimated pIC_{50} error available for each predicted compound.

Overall, approaches based on the SVR algorithm and Dragon descriptors produced the best results. Within this study, Dragon descriptor-based models performed best for all prediction methods. However, the gain of predictive performance through additional 2000 descriptors was not as high as one could expect. MOE descriptor-based analogues were nearly equally predictive. While on the one hand SVR models performed best, on the other hand, as mentioned above, their prediction performance strongly depended on the tuning parameters chosen for these models. In contrast, the Random Forest algorithm is very robust with regard to different tuning parameters. Compared to SVR, RF is relatively easy to use and performs well even with default settings. Within this case study, it became clear that PLS is not suitable to handle such noisy data. However, we have been able to show that SVR and RF algorithms yield good results in our generalized workflow for generating predictive in silico off-target activity models using heterogeneous in vitro data from bioactivity databases. Moreover, this standardized workflow is easily transferable to other off-targets and assay readouts.

CONCLUSIONS AND OUTLOOK

Chemical structure data and measured bioactivities of compounds are nowadays easily available from public and commercial databases. Usage of such databases for modeling purposes in academia[62] and industry[2,9] has substantially increased over the last years. However, these databases contain heterogeneous data coming from different laboratories determined under different protocols and, in addition, sometimes even erroneous entries.[18,19,34] At the same time, much research has focused on toxicity prediction in early drug discovery and crop-protection research. In silico screening for classic toxicological end points like genotoxicity for example with the help of structural alerts is already quite common.[3,63−65] In this study, we evaluated the use of heterogeneous in vitro data from bioactivity databases covering more than 2,200 chemical structures for the development of predictive QSAR models for off-target mediated toxicity. Human acetylcholinesterase inhibition was chosen as a data-rich and toxicologically significant end point for our case study.

The presented data preparation workflow forms the fundamental basis of this work. This standardized and thorough quality management routine for inhomogeneous input data finally enables the development of predictive QSAR models based on in vitro data from multiple laboratories. An extended applicability domain approach was used, and regression results were refined by an error estimation routine. The obtained in silico models for human AChE inhibitory potential show excellent accuracy in external validation (see Table 3). These results were achieved through detection and special consideration of borderline candidates. However, the current setup of our generalized workflow is not suited to assess the respective influence or contribution of individual molecular descriptors. Future work will be devoted to addressing this issue.

The modular concept of the approach makes it easy to implement other machine learning techniques and molecular descriptors. To further increase a models' predictivity or to increase the chemical space covered by prediction models, new data can be easily added to the preparation workflow. Thus, the whole standardized process is easily transferable to other off-targets and assay readouts.

We propose this procedure as a standard approach for generating in silico models for off-target mediated toxicity. Off-target related early toxicity predictions address important aspects of potential compound failures. Furthermore, we regard this approach as a contribution to the goal of reducing experiments on animals in the course of the "3R" concept (Replacement, Reduction, and Refinement). The developed models are well-suited as screening tools for toxic potential in pharmaceutical or crop protection research in order to prioritize chemical classes for further development or experimental testing in early research phases.

ASSOCIATED CONTENT

Ⓢ Supporting Information

IDs of extracted data from bioactivity databases ChEMBL and WOMBAT, tuning parameters for combinations of machine learning algorithms and descriptor sets, list of blacklisted data points containing errors in extracted data, and full list of results for all tested tuning parameter combinations. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: klaus-juergen.schleifer@basf.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank our colleagues Ian Craig, Martin Weisel, Eric Fabian, and Hennicke Kamp for their critical proof reading and fruitful discussions.

■ REFERENCES

(1) Taniguchi, C. M.; Armstrong, S. R.; Green, L. C.; Golan, D. E.; Tashjian, A. H. Drug Toxicity. In *Prinicples of Pharmacology: The Pathophysiologic Basis of Drug Therapy*, 2nd ed.; Golan, D. E., Tashjian, A. H., Armstrong, E. J., Armstrong, A. W., Eds.; Lippincott Williams & Wilkins: Baltimore, MD, 2011; pp 63−73.

(2) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J. L.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **2012**, *486*, 361−367.

(3) Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Mueller, L.; Paehler, A. Computational toxicology in drug development. *Drug Discovery Today* **2008**, *13*, 303−310.

(4) Lahl, U.; Gundert-Remy, U. The Use of (Q)SAR Methods in the Context of REACH. *Toxicol. Mech. Methods* **2008**, *18*, 149−158.

(5) Bennett, P. B.; Guthrie, H. R. E. Trends in ion channel drug discovery: advances in screening technologies. *Trends Biotechnol.* **2003**, *21*, 563−569.

(6) Benigni, R.; Netzeva, T. I.; Benfenati, E.; Bossa, C.; Franke, R.; Helma, C.; Hulzebos, E.; Marchant, C.; Richard, A.; Woo, Y. T.; Yang, C. The Expanding Role of Predictive Toxicology: An Update on the (Q)SAR Models for Mutagens and Carcinogens. *J. Environ. Sci. Health., Part C* **2007**, *25*, 53−97.

(7) Matter, H.; Anger, L. T.; Giegerich, C.; Guessregen, S.; Hessler, G.; Baringhaus, K.-H. Development of in silico filters to predict activation of the pregnane X receptor (PXR) by structurally diverse drug-like molecules. *Bioorg. Med. Chem.* **2012**, *20*, 5352−5365.

(8) Thai, K.-M.; Ecker, G. F. Predictive models for hERG channel blockers: Ligand-based and structure-based approaches. *Curr. Med. Chem.* **2007**, *14*, 3003−3026.

(9) Czodrowski, P. hERG Me Out. *J. Chem. Inf. Model.* **2013**, *53*, 2240−2251.

(10) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20*, 241−266.

(11) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct. (Theochem.)* **2003**, *622*, 39−51.

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

(13) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed January 13, 2014).

(14) ChEMBL_15. http://www.ebi.ac.uk/chembl/ (accessed July 10, 2013).

(15) DrugBank. http://www.drugbank.ca/ (accessed January 13, 2014).

(16) Liceptor. http://www.evolvus.com/ (accessed January 13, 2014).

(17) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Wiley-VCH Verlag GmbH & Co. KGaA: 2005; pp 223−239.

(18) Tiikkainen, P.; Franke, L. Analysis of Commercial and Public Bioactivity Databases. *J. Chem. Inf. Model.* **2012**, *52*, 319−326.

(19) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC50 Data - A Statistical Analysis. *PLoS One* **2013**, *8*, e61007.

(20) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494−3504.

(21) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **2010**, *29*, 476−488.

(22) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.

(23) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570−2578.

(24) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.

(25) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W. D.; Veith, G.; Yang, C. H. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52. *ATLA Altern. Lab. Anim.* **2005**, *33*, 155−173.

(26) Sushko, I.; Novotarskyi, S.; Koerner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Mueller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Oeberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094−2111.

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

(28) Soreq, H.; Seidman, S. Acetylcholinesterase - new roles for an old actor. *Nat. Rev. Neurosci.* **2001**, *2*, 294−302.

(29) Grob, D.; Harvey, J. C. Effects in Man of the Anticholinesterase Compound Sarin (Isopropyl methyl phosphonofluoridate). *J. Clin. Investig.* **1958**, *37*, 350−368.

(30) Kamanyire, R.; Karalliedde, L. Organophosphate toxicity and occupational exposure. *Occup. Med.-Oxf.* **2004**, *54*, 69−75.

(31) Ellman, G. L.; Courtney, K. D.; jr, V. A.; Featherstone, R. M. A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochem. Pharmacol.* **1961**, *7*, 88−95.

(32) Sharma, A.; Piplani, P. Acetylcholinesterase Inhibitors from QSAR Point of View: How Close are We? *Cent. Nerv. Syst. Agents Med. Chem.* **2013**, *13*, 71−87.

(33) Yan, A.; Wang, K. Quantitative structure and bioactivity relationship study on human acetylcholinesterase inhibitors. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 3336−3342.

(34) Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mractc, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology*; Wiley-VCH Verlag GmbH: 2008; pp 760−786.

(35) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(36) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.

(37) *The Open Babel Package* version 2.3.2. http://openbabel.org/ (accessed May 11, 2013).

(38) The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi/ (accessed May 11, 2013).

(39) O'Boyle, N. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **2012**, *4*, 22.

(40) Mosteller, F.; Tukey, J. W. Data analysis, including statistics. In *The Handbook of Social Psychology*, 2nd ed.; Lindzey, G., Aronson, E., Eds.; Addison-Wesley: Reading, MA, 1968; Vol. 2, pp 80−203.

(41) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Koetter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. In *KNIME: The*

Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), 2007; Springer: 2007.

(42) R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2013.

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

(44) *DRAGON (Software for Molecular Descriptor Calculation)*, version 6.0; Talete srl 2013.

(45) *Molecular Operating Environment (MOE)*, version 2012.10; Chemical Computing Group Inc.:Montreal, QC, Canada, 2012.

(46) Halgren, T. A. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem.* **1996**, *17*, 616−641.

(47) Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *2*, 559−572.

(48) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417−441.

(49) Wold, S.; Sjoestroem, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

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

(51) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.

(52) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(53) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(54) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, *49*, 2481−2488.

(55) Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; Zhu, H.; Rusyn, I.; Tropsha, A. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chem. Res. Toxicol.* **2011**, *24*, 1251−1262.

(56) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Rev. Comput. Chem.*; John Wiley & Sons, Inc.: 2007; pp 291−400.

(57) Li, H.; Liang, Y.; Xu, Q. Support vector machines and its applications in chemistry. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 188−198.

(58) Cooper, J. A.; Saracci, R.; Cole, P. Describing the Validity of Carcinogen Screening-Tests. *Br. J. Cancer* **1979**, *39*, 87−89.

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

(60) Tropsha, A. Potential of short-term biological assays to quantitatively predict chronic toxicity. *Toxicol. Lett.* **2013**, *221*, S52−S53.

(61) Belluti, F.; Rampa, A.; Piazzi, L.; Bisi, A.; Gobbi, S.; Bartolini, M.; Andrisano, V.; Cavalli, A.; Recanatini, M.; Valenti, P. Cholinesterase Inhibitors: Xanthostigmine Derivatives Blocking the Acetylcholinesterase-Induced beta-Amyloid Aggregation. *J. Med. Chem.* **2005**, *48*, 4444−4456.

(62) Tsareva, D. A.; Ecker, G. F. How Far Could We Go with Open Data - A Case Study for TRPV1 Antagonists. *Mol. Inf.* **2013**, *32*, 555−562.

(63) Ashby, J.; Tennant, R. W. Chemical-structure, Salmonella Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis among 222 Chemicals tested in Rodents by the United-States NCI/NTP. *Mutat. Res.* **1988**, *204*, 17−115.

(64) Hillebrecht, A.; Muster, W.; Brigo, A.; Kansy, M.; Weiser, T.; Singer, T. Comparative Evaluation of in Silico Systems for Ames Test Mutagenicity Prediction: Scope & Limitations. *Chem. Res. Toxicol.* **2011**, *24*, 843−854.

(65) Sanderson, D. M.; Earnshaw, C. G. Computer-Prediction of possible toxic action from Chemical-Structure - the DEREK system. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.