

## RESEARCH ARTICLE

### Prediction of chemical toxicity with local support vector regression and activity-specific kernels

Andreas Maunz<sup>a,†</sup> and Christoph Helma<sup>a,b,‡</sup>

<sup>a</sup>*Freiburg Center for Data Analysis and Modelling (FDM), Hermann-Herder-Str. 3a, 79104 Freiburg, Germany;* <sup>b</sup>*in silico toxicology, Talstr. 20, 79102 Freiburg, Germany*

*(v1.0 released February 2008)*

---

<sup>†</sup>Corresponding author. Email: maunza@fdm.uni-freiburg.de, phone: +497612035807, fax: +497612037700

<sup>‡</sup>Email: helma@in-silico.de, phone: +491707591929, fax: +497612037700

We propose a new kernel, based on 2-D structural chemical similarity, that integrates activity-specific information from the training data, and a new approach to applicability domain estimation that takes feature significances and activity distributions into consideration. The new kernel provides superior results than the well-established Tanimoto kernel, and activity-sensitive feature selection enhances prediction quality. Validation of local support vector regression models based on this kernel has been performed with three publicly available datasets from the DSSTox project. One of them (Fathead Minnow Acute Toxicity) has been already modelled by other groups, and serves as a benchmark dataset, the other two (Maximum Recommended Therapeutic Dose, IRIS Lifetime Cancer Risk) have been modelled for the first time according to the knowledge of the authors. For all three models predictive accuracies increase with the prediction confidences that indicate the applicability domain. Depending on the confidence cutoff for acceptable predictions we were able to achieve > 90% predictions within 1 log unit of the experimental data for all datasets.

**Keywords:** Structure kernels; Instance-based models; Lazy Learning; Fathead Minnow Acute Toxicity; Maximum Recommended Therapeutic Dose; IRIS Lifetime Cancer Risk

## 1. Introduction

The use of *in-silico* methods to assess the toxic potential of chemicals has increased steadily in recent years. They are used for preliminary testing and the prioritization of chemicals, and are supposed to have the potential to reduce animal experiments.

Traditional (Q)SAR techniques assume a common biological mechanism (or *mode of action*, MoA [1]), but most existing toxicological datasets (e.g. PubChem [<http://pubchem.ncbi.nlm.nih.gov/>], Toxnet [<http://toxnet.nlm.nih.gov/>], DSSTox [<http://www.epa.gov/nheerl/dsstox/>]) are structurally very diverse. As toxicological experiments are usually expensive, time consuming, and may require a large number of experimental animals it is frequently impossible to create experimental data for congeneric compounds specifically for (Q)SAR modelling.

To circumvent this problem compounds can be divided into subsets that are supposed to exhibit a single mode of action (MoA), and to create separate (Q)SAR models for these subsets. Chemical similarity, chemical classes, and the presence of structural alerts are used as selection criteria, but they are frequently ambiguous and do not necessarily guarantee a common mode of action.

Modern machine learning techniques (e.g. neural networks, support vector machines) allow much more expressive models than linear regression, but they are hard to interpret in terms of toxicological mechanisms, and may suffer from data sparseness in high dimensions.

As an alternative, we propose to detect clusters of similar compounds automatically, and to generate local regression models for each cluster. For efficiency reasons, it is advisable to avoid to create models for all possible similarity clusters, but to defer model building until a query compound is known (instance-based models, also termed *lazy learning* [2]). This means that for each query compound a specific local (Q)SAR model is created, that relies only on structurally similar compounds from the training set. It is possible to interpret such a procedure as automated categorization, read across, and analogue detection with the aim to introduce sound statistical criteria instead of manual intervention.

Several instance-based models for chemical predictions have been proposed, including nearest-neighbor prediction for classification [3], weighted average, and locally weighted regression [4, 5]. In this paper we present a novel technique based on local support vector regression with activity specific kernels, and evaluate them on three toxicological datasets (acute fathead minnow toxicity, maximum recommended therapeutic dose, IRIS lifetime cancer risk) from the DSSTox project.

The remainder of this article is organized as follows: First, we introduce a novel kernel function based on the significance of structural features (the significance-weighted kernel), provide a formal definition of applicability domains, and illustrate with an example how to interpret a prediction. Second, we compare for a well-known dataset (fathead minnow toxicity) the significance-weighted kernel with the widely used Tanimoto kernel [8]. Third, we limit the use of structural features to the most significant ones, and compare the validation results for fathead minnow toxicity and two previously unpredicted datasets (maximum recommended therapeutic dose, lifetime cancer risk). Finally, we evaluate the resulting models according to the OECD principles.

## 2. Methods

An introduction to kernel methods would be out of scope of this paper, so the reader is assumed to have basic knowledge in this area [6]. Intuitively, a kernel defines a measure of similarity between two instances (here: molecules). In a comparative study [7], different spectral kernels based on chemical structure were defined and evaluated, and we largely follow the conventions and formal notations employed there. Among the kernels studied, the Tanimoto kernel was evaluated favorably. The related Tanimoto index is one of the most useful chemical similarity indices, as shown in another study [8]. In this paper we introduce the significance-weighted kernel, that takes the significance of fragments for a particular endpoint into account. It differs from

the standard Tanimoto kernel by assigning different weights to features rather than considering their mere presence. Therefore, the new kernel describes activity-related similarity, instead of general structural similarity as the Tanimoto kernel.

A linear fragment of a compound is a linear subgraph (or path) of the 2D graph representing the compound. Linear fragments are allowed to share edges, but cycles are not allowed. For this study, the features were obtained with depth-first search, using a simplified version of MOLFEA, the Molecular Feature Miner [9], employing the SMILES representation language [10]. The depth of the search was not constrained in length, which could also be added to limit computational effort. Otherwise, due to the small branching factor for small molecules, such features can be efficiently calculated [7].

The training data is inherently informative about the significance of features. The detection of significant features associates a significance value  $p_f$ , also referred to as  $p$ -value, with every feature  $f$  in the training set (i.e. the union of all features from all training compounds). The Kolmogorov-Smirnov test (KS test) is used to identify features that correlate with the endpoint under consideration<sup>1</sup> [11]. The KS test compares two cumulative probability distributions sampled from quantitative data, and returns a  $p$ -value indicating the probability that the two sets were drawn from the same probability distribution (null hypothesis). An example is shown in Figure 1.

FIGURE 1

In our case, the KS test is used to assess whether the activity values  $A_{|f}$  for training compounds containing  $f$  differ significantly from the activity values of all training compounds  $A$ , whether they contain  $f$  or not<sup>2</sup>. The  $p$  values are converted to  $1 - p$ , because low  $p$  values yield the most significant features. In summary, for every feature  $f$ , we have an associated significance value  $p_f$ , that is obtained from  $A$  and  $A_{|f}$ . At least two important properties of features can be extracted from this statistical analysis:

- A feature is called significant if its  $p$ -value is  $\geq 0.9$ .
- A feature is called activating, if the median of  $A_{|f}$  is lower than the median of  $A$ , and deactivating, if it is higher<sup>3</sup>.

The above procedure works for all features present in the training database. However, it is possible that a fragment is present in the query structure but not in the training set. Such a fragment is called unknown and no  $p$ -value can be calculated for such fragments.

## 2.1 Significance-weighted kernel

We can now define fingerprint similarities between each two compounds. It should be noted, that two cases can be distinguished:

- One is the query structure and the other one is a training compound.
- Both are training compounds.

In the first case, the similarity is additionally multiplied by the fraction of known fragments, i.e. a penalization for unknown fragments is implemented. This is especially important for small training sets (see also validation of the IRIS dataset, sections 3.2 and 3.5).

Spectral graph theory is concerned with the relationship between the adjacency matrix of a graph and its characteristic polynomial, eigenvalues, and eigenvectors. For this spectral kernel application, the training compounds are regarded as nodes of a fully connected graph. A weight

<sup>1</sup>For qualitative activities (classification), the chi-square test can be used.

<sup>2</sup>The sample size for the test was not restricted.

<sup>3</sup>Activity is commonly measured in quantities needed to obtain a toxicological effect, so higher values mean lower activity.

equal to the kernel value between two adjacent nodes is assigned to each edge of the graph, enabling similarity weighting. The adjacency matrix of the graph is equivalent to the kernel matrix, also known as the Gram matrix.

Let  $\mathcal{F}$  be the ordered set  $\{f_1, \dots, f_m\}$  of all possible linear fragments of size  $|\mathcal{F}| = m$ . Each molecule  $\mathbf{u}$  can be compactly represented as a real valued vector of length  $m$ , where the  $i$ -th position equals  $p_{f_i}$ , if  $f_i$  is contained in  $\mathbf{u}$ , and 0 else, for  $i \in \{1 \dots m\}$ <sup>4</sup>. It is possible to reduce this real-valued vector to a binary vector (up to precision  $\frac{1}{10^r}, r \in \mathbb{N}$ ) by representing every position  $i$  as binary vector  $\in \{0, 1\}^{10^r}$  by setting the first  $\lfloor p_{f_i} * 10^r \rfloor$  bits to 1, and the rest to 0, yielding a binary weight map  $\phi^w(\mathbf{u})$  of length  $m * 10^r$ .

The simple dot-product on these fingerprints is a measurement of activity-specific similarity between each two compounds  $\mathbf{u}, \mathbf{v}$ , i.e.

$$k(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^{m*10^r} \phi^w(\mathbf{u})_i \phi^w(\mathbf{v})_i. \quad (1)$$

Note, that if the compounds are equal, then  $k(\mathbf{u}, \mathbf{v}) = m * 10^r$ . Conversely, if they do not share any fragments,  $k(\mathbf{u}, \mathbf{v}) = 0$ . The more significant a fragment, the greater its weight. Moreover,  $k(\mathbf{u}, \mathbf{v})$  is the sum of (the  $p$ -values of) the features in the intersection between  $\mathbf{u}$  and  $\mathbf{v}$ . In order to standardize this value, we relate it to the sum corresponding to the union of features between  $\mathbf{u}$  and  $\mathbf{v}$ , which can be obtained by  $k(\mathbf{u}, \mathbf{u}) + k(\mathbf{v}, \mathbf{v}) - k(\mathbf{u}, \mathbf{v})$ . The significance-weighted kernel is then defined as

$$k^w(\mathbf{u}, \mathbf{v}) = \frac{k(\mathbf{u}, \mathbf{v})}{k(\mathbf{u}, \mathbf{u}) + k(\mathbf{v}, \mathbf{v}) - k(\mathbf{u}, \mathbf{v})} = \frac{k(\mathbf{u}, \mathbf{v})}{2m * 10^r - k(\mathbf{u}, \mathbf{v})}. \quad (2)$$

Note, that if all  $p$ -values are set to 1, this equals the standard Tanimoto kernel. A proof sketch that the significance-weighted kernel is a positive definite kernel (*Mercer kernel*) follows from a result by Gower [12], showing that, for any integer  $p$  and any set  $l$  of binary vectors,  $\mathbf{x}_1, \dots, \mathbf{x}_l \in \mathbb{R}^p$ , the similarity matrix  $S = k^w(\mathbf{x}_i, \mathbf{x}_j)_{1 \leq i, j \leq l}$  is positive semidefinite. Thus,  $k^w$  is a positive definite kernel [7].

In summary, the  $p$ -values for fragments enrich the kernel's expressiveness with activity-specific information. In consequence, similarity is not only based on structural features, but also on activity, and the neighbors are hence similar with regard to the current endpoint.

## 2.2 Local Regression Models

As most toxicological datasets contain results for *non-congeneric* compounds, model building is deferred until the query structure is known. Only compounds that are similar to the query structure are used to build a local regression model. To determine structural similarity with respect to a particular endpoint, the weighted Tanimoto index (see previous section) is used.

Only compounds  $\mathbf{x}_i$  with  $k^w(\mathbf{x}_q, \mathbf{x}_i) > 0.3$  are considered neighbors (referred to as the set  $N$ ) to the query structure. If no neighbors for a query structure can be found according to this definition, no prediction is made. Preliminary trials showed that using a lower threshold yielded no significant information gain. Higher values, on the other hand, led to excessive amounts of empty neighborhoods and therefore missing predictions.

Local regression models are trained with  $\nu$ -sv regression [13]. The parameter  $\nu$  was set to 0.8 which turned out to constitute the best overall performance as indicated by preliminary trials.

---

<sup>4</sup>Unknown fragments are also weighted with 0.

### 2.3 Example Prediction

Figure 2 illustrates the practical consequences of our algorithms with an example from the EPAFHM and FDAMDD datasets (see section 2.5). It shows the two most similar neighbors and the most significant fragments for the query structure with the SMILES code CNCCCNC3c1cccc1CCc2cccc23 (Desipramine).

FIGURE 2

Notice, that neighbor 1 of the FDAMDD prediction (Trimipramine) has a similarity of 1.0, although the two structures are different. They rather share the same *significant* features. Likewise, similarities (and as a result neighbors) depend on the selected endpoint as well. Therefore, for two different training databases, and consequently two different endpoints, the same pair of compounds does not necessarily have the same similarity. Accordingly, the 10 most significant fragments, depicted in the lower part of Figure 2, differ for the two training sets.

### 2.4 Applicability Domain Estimation

Due to chemical diversity, every (Q)SAR model has only a limited domain of applicability, namely “the physico-chemical, structural or biological space on which it has been trained” [14]. Consequently, it can only make valid predictions for this domain.

It has been shown, that training compounds more similar to the query structure give better predictions [15]. In our approach, a neighborhood similar to the query compound with respect to structure and activity is automatically mined from the training data. The process of finding neighbors generates valuable information that can be incorporated into a confidence index indicating the reliability of the prediction, considering both dependent and independent variables [14].

- The higher the median similarity  $\tilde{s}$  of the neighbors, the higher the confidence in the prediction<sup>5</sup>.
- Conversely, the higher the standard deviation  $\sigma_a$  of activity values of the neighbors, the lower the confidence.

A variety of kernel functions has been reviewed for smoothing similarities [2]. In the actual implementation of the confidence index the gaussian smoothed median similarity  $\tilde{s}$  was assessed by smoothing the single neighbor similarities, and taking the median of these values:

$$\tilde{s} = \text{median}\{ \varphi_{0.3} ( 1 - k^w(\mathbf{x}_q, \mathbf{x}_i) ) \mid \forall \mathbf{x}_i \in N \}, \quad (3)$$

where  $\varphi_\sigma(x) = e^{-\frac{x^2}{2\sigma^2}}$  is the gaussian squared exponential function with standard deviation  $\sigma$ . The confidence value was defined as

$$\text{conf} = \tilde{s}e^{-\sigma_a}. \quad (4)$$

The exponential rapidly “punishes” large  $\sigma_a$  but stays close to 1 for small values. This was introduced due to the fact that the biggest errors occurred for the most active compounds. Closer investigation revealed greatly varying activity values of the neighbors. This is plausible because those compounds require the smallest doses for a reaction, and measurement errors were more likely in the experiments that generated the data. In summary, the confidence index indicates the structural “density” and similarity in activity for the neighborhood, derived by

<sup>5</sup>The median is used rather than the mean to reduce the sensitivity to eventual skew in the data.

statistical measures, which –of course– don't imply causality for a particular MoA. However, it describes prediction conditions pretty well (see next sections).

## 2.5 Data

Three publicly available diverse datasets were used for validation:

- the EPAFHM dataset (Fathead Minnow Acute Toxicity), specifically the  $LC_{50}$  values (`lc50_mmol`, 573 compounds) from `EPAFHM_v4a_617_15Jun2007.sdf` [16]. This dataset was explicitly designed for (Q)SAR studies and contains mode of action (MoA) classifications for the chemicals. However, this information was not used, since the goal was to find similar training sets automatically. The activities span 9.34 orders of magnitude.
- the FDAMDD dataset (Maximum Recommended Therapeutic Dose based on clinical trial data), specifically the maximum recommended therapeutic dose (`mmol/kg-bw/day`, `dose_mrdd_mmol`) from `FDAMDD_v3a_1216_25Jul2007.sdf` ([17], 1215 pharmaceutical compounds). The compounds were used for the oral treatment of patients, usually 3-12 months. The activities span 8.93 orders of magnitude.
- part of the IRIS dataset (Human health effects due to exposure to various substances found in the environment), specifically the upper-bound excess lifetime cancer risk from continuous exposure to  $1 \mu\text{g/L}$  in drinking water (`drinkingwater_unitrisk_micromol_per_l`, 68 compounds) from `IRISTR_v1a_544_28Jul2007.sdf` [18]. The activities span 5.77 orders of magnitude.

The first dataset has been used in a lot of studies, some of the more recent ones are reported here for comparison [19–22]. To the knowledge of the authors, the latter two datasets have not yet been validated with (Q)SAR models. To ensure a more normal distribution, all activity values were transformed to log values. For all three datasets all compounds were used, i.e. no compounds were excluded *a priori*.

## 2.6 Validation

For validation, data sets were split into folds of size 1 and 10% to perform leave-one-out crossvalidation (LOO-CV) and 10-fold crossvalidation (10-CV). This means that we repeatedly removed 1 compound or 10% of the compounds from the training data and predicted them on the basis of the rest of the dataset. All crossvalidation techniques were performed in a way that prohibits information flow from the test set into the training set. In particular

- the selection of features was repeated for each fold, and
- $p$ -values were recalculated from scratch for each fold.

Previous investigations demonstrated theoretically and empirically that LOO-CV provides a good estimation of predictive accuracies, if the applicability domain is considered [23]. A confidence value was assigned to each prediction (see section 2.4). Different predictivity measures were assessed in a cumulative fashion in descending confidence order, meaning that, for every confidence level, the number describes the predictions with higher or equal confidence.

- $nr$  is the number of predictions made.
- $r^2$  is the multiple correlation coefficient obtained by internal validation.
- $q^2$  is the cross-validated coefficient that indicates the explained variance (only reported for set sizes  $\geq 12$  due to missing statistical reliability for smaller set sizes [24]).
- *Weighted accuracy* ( $wa$ ) is the fraction of predictions within the 1 log unit error margin [25], weighted by their prediction confidence. Confidence weighting shall ensure that predictions with high confidences have a higher impact on validation results than predictions with low confidence.
- *Mean error* ( $me$ ) is the mean of the raw prediction errors, i.e. the errors are not standardized.

- *RMSE* (*rmse*) is the root mean-squared error of the predictions.

## 2.7 Implementation

Feature calculation was implemented with the `OpenBabel` C++ library [26, 27] and the `kernlab` package for R [13, 28] was used for regression. A web based graphical user interface was developed with the "Ruby on Rails" Model-View-Controller framework.

Source code for this project and installation instructions for Linux are available from `svn://www.in-silico.de/lazar/branches/dist`. A web based graphical user interface can be obtained from `svn://www.in-silico.de/opentox/lazar-gui`.

Prediction models for the endpoints described in this article will be available for the general public at the website `http://lazar.in-silico.de`.

## 3. Results and Discussion

### 3.1 Significance-weighted vs. unweighted Tanimoto kernel

The predictive performance of the significance-weighted kernel and the standard Tanimoto kernel were compared for the EPAFHM dataset. We expected for both kernels increasing performance with increasing confidence thresholds, with advantages for the significance-weighted kernel. The LOO-CV data (see Table 1) gives performance measures for both kernels and different confidence thresholds. All features were used in the runs, i.e. there was no *p*-value threshold applied. The  $q^2$  values are plotted in Figure 3. Clearly, the significance-weighted kernel outperforms the unweighted Tanimoto kernel in *nr*,  $q^2$ , *me* and *rmse*, indicating that the *p*-values are indeed meaningful, and can help to identify features that are significant for the exhibited activity. We suggest that the better *wa* values for the Tanimoto kernel may be an effect of the lower number of predictions made by this kernel (as discussed below).

FIGURE 3

Further LOO-CV runs were conducted to assess the effects of the *p*-values more closely. For that purpose, a significance-threshold was introduced for *p*-values, meaning that, for every run, only features with *p*-values equal or higher to the threshold were used in the kernel function. The step-width was 0.1 and the values ranged between 0.0 and 0.9. Regarding median values across the same confidence thresholds as above, we expected the advantage for the significance-weighted kernel to shrink, because leaving out features with low *p*-values can also be considered a (extreme) form of weighting, which should raise the performance of the standard Tanimoto kernel to (nearly) the level of the significance-weighted kernel. Table 2 gives performance measures for the different significance thresholds and Figure 4 plots values for *nr* and *rmse*. For low thresholds (0.0-0.2), the standard Tanimoto kernel produces much larger *rmse* values. Above 0.3, both kernels have similar performance<sup>6</sup>. Note, however, that in general performance still increases for higher thresholds.

FIGURE 4

Interestingly, for a broad range of low thresholds (0.0-0.5) the significance-weighted kernel is able to predict much more compounds while the number of predictions decreases for both kernels and high thresholds (0.8-0.9). The first finding is due to the high weight that features with low *p*-values get in the unweighted kernel (recall that the kernel values indicate similarity and are

<sup>6</sup>The *rmse* and  $q^2$  value of 0.92 at significance threshold 0.5 for the significance-weighted kernel results from a single severely mispredicted data point that did not occur for the other values. Leaving it out gives a value of 0.74

defined as the fraction of common features, see section 2.1). This seems to massively decrease similarity which means that they are widely distributed. The second finding is probably due to sparseness and also wide distribution of features with high  $p$ -values.

### 3.2 Predictive Model Validation

The following reports predictive validation results for all three datasets for various confidence thresholds, depicted in detail in Figure 5. Table 4 summarizes the data in the form of median values across the confidence threshold (a detailed tabulated version, as well as scatterplots of predicted vs database activities, are available from the website of this article as supporting information).

For the FDAMDD and EPAFHM dataset, LOO-CV and 10-CV results are given. For the IRIS dataset, due to the small set size, only LOO-CV results are presented. Additionally, for the IRIS dataset, Figure 5 gives only results for the use of significant features. For all datasets,  $q^2$  values are only presented for sufficient set sizes (i.e.  $\geq 12$ ).

FIGURE 5

The number of predictions is normally distributed in the confidence range, as can be seen by the curve for  $nr$ , which is close to a cumulative gaussian distribution function for all datasets. Confidence values (that indicate the applicability domain) have two major effects on prediction quality.

- The  $wa$  and  $r^2/q^2$  values increase for increasing confidence thresholds. Thus, the fraction of predictions outside the 1 log unit error margin decreases while the linear fit increases.
- Consequently, the  $rmse$  values decrease for increasing confidence thresholds.

These effects are most prominently visible for FDAMDD, the largest dataset (see Figure 5c and d). For the EPAFHM dataset, the  $rmse$  values do not decrease monotonically but  $q^2$  values increase monotonically. This is due to a smaller “cloud” around the center of the distribution for higher confidence values. For the IRIS dataset,  $rmse$  and  $q^2$  values behave again monotonically. The 10-CV numbers follow largely the LOO-CV values, differences are the consequence of the smaller training set sizes.

FIGURE 6

Response permutation testing [29] was used to assess the significance of the obtained  $q^2$  values for the EPAFHM and IRIS dataset. This method compares  $q^2$  values obtained with randomly permuted training activities to the  $q^2$  value obtained with the correct activities. Here, 50 LOO-CV runs for a confidence threshold of 0.225 were used to sample such a  $q^2$ -distribution (see Figure 6). The EPAFHM results clearly indicate a significant model, ruling out the possibility of chance correlation between predictions and database activities. For the IRIS dataset, the  $q^2$  values vary more, but only 4 of them are positive. The intercept is clearly below 0.05 in both cases, indicating a valid model [29].

Two aspects of the validation results in Figure 5 and Table 4 are most prominent:

- The exclusive use of significant features increases accuracy significantly for *high* confidence predictions. The most impressive gain was obtained for the IRIS dataset where the median  $rmse$  values could be halved, from 1.07 to 0.56, and  $q^2$  values increased from below 40% to nearly 80%. For the EPAFHM dataset, up to 10% of  $q^2$  gain could be achieved, and for the FDAMDD dataset accuracy increased rather slightly.
- However, using all features significantly increases accuracy for the *lower* confidence values. For the EPAFHM dataset, for instance, the best 147 predictions achieve  $q^2$  value of 0.74 compared to 0.62 when using significant features only. Also, the numbers of high and low

confidence predictions are much more equally distributed, as can be seen from the *nr* median values.

We suspect that the reasons for these findings lie in the sparse distribution of significant features that are, however, highly informative. If present, they enable high accuracies (i.e. in high confidence predictions), but non-significant features can compensate partially in their absence.

In a broader sense, the use of significant features only vs. using all features can be seen as a switch between two modes: the former enables high-precision mode for the top-confidence predictions, the latter increases the coverage of the dataset with still reasonable performance. For instance, the median value of *nr* (number of predictions) is more than doubled from 76 to 165, indicating much more data in the low-confidence part, when using significant features only. Also more data is predicted in total when using all features (568 compounds compared to 529). These findings for the EPAFHM dataset also hold for the other datasets.

The integration of the ratio of known fragments into the kernel function plays an important role, because it indicates the presence of substructures that cannot be evaluated with the training set. To evaluate the impact of this factor, LOO-CV was performed without compensation for unknown fragments. For the two large datasets EPAFHM and FDAMDD, only minor effects were observed, but for the IRIS dataset,  $q^2$  values dropped rapidly from over 80% to about 60%. We assume this is due to infrequent occurrences, and therefore generally large “gaps”, characteristically for small datasets.

### 3.3 Fathead Minnow Toxicity

The EPAFHM dataset is the only dataset that has been used extensively for QSAR modelling.

Presently available models [19–22] used global modelling techniques with fixed training and test sets, and three of them made use of more or less heavily preselected similar chemicals for training and/or test sets (according to MoA or other criteria, e.g. structural dissimilarity expressed in leverage values). None of these studies used the original EPAFHM set, but rather left out compounds, and/or integrated data from other sources. The criteria for such selections are frequently unclear, which allows us to draw only limited conclusions from a comparison.

Papa et al. [19] used a selected set of 468 compounds from the EPAFHM dataset, and excluded further 19 structurally heterogeneous outliers that strongly affected the performance of their model. They estimated the applicability domain with the leverage approach. Öberg [21] modelled 311 narcotics from the EPAFHM dataset, selected on the basis of narcotics mode I-III, split 33% as test set, and used the rest as training set. Niculescu et al. [20] used the EPAFHM dataset and added more compounds (total 886). They used a test set of 86 compounds, but excluded 20 compounds that could not be fit by their neural network. Pavan et al. [22] collected 408 diverse chemicals from different sources (among them the EPAFHM dataset) as training set, and predicted 57 compounds from a different source (external prediction). They estimated the applicability domain with the leverage approach. The reported results of those four studies are given in the upper part of Table 3.

For comparison purposes, we created a test set by randomly removing 20% of instances of the dataset (114 compounds), and using the rest as training set. Since applicability domain estimation is integrated into the algorithm in the form of real-valued confidence estimates (as described in section 2.4), values for different confidence thresholds are reported. For reference, we also report results for another one of our models, a local multilinear approach, that builds directly on the linear fragments as descriptors. It employs objective feature selection in conjunction with principle components analysis. The corresponding figures for different confidence thresholds are given in the lower part of Table 3. In summary, 72% of the test set can be predicted with good accuracy using automated selection methods, yielding a reasonable coverage and generalization performance. This seems to be comparable with models from the literature, especially under the aspect that no “outliers” were removed, and no expert information (e.g. about mode of actions) was required as input.

### 3.4 *Maximum Recommended Therapeutic Dose*

Out of the three examined datasets, FDAMDD was the largest, and, according to the knowledge of the authors, no other (Q)SAR models exist for this endpoint. According to the DSSTOX project, most of the values in this dataset were determined from pharmaceutical clinical trials using oral daily treatments, and a wide range of adverse or toxic effects were considered in assigning the MRTD. Considering that it spans nearly 9 orders of magnitude of dose variation, the prediction of 773 compounds (70%) with  $q^2 = 64\%$  is a good coverage. The use of significant features still increased the performance of our algorithm for this dataset. Although the publishers of the dataset state that in many cases only a single derivative of a known family of drugs is represented in the database, our structural kernel method can deal well with its chemical diversity. The good performance of this model is also an indication that it is indeed possible to predict human endpoints reliably with QSAR techniques.

### 3.5 *IRIS Lifetime Cancer Risk*

The tiny IRIS dataset consists of expert estimations of lifetime cancer risk. The information in IRIS is intended for use in protecting public health through risk assessment and risk management. Also here, to the knowledge of the authors, no other (Q)SAR models exist for this endpoint. Originally, it was supposed to be “of limited use for direct structure-activity relationship (SAR) modeling due to the many extrapolations and judgements applied to the objective test data” [18]. Despite this fact and its small size our models show good prediction performance when using significant features. This indicates that our technique is not only capable to model experimental data, but also to approximate human expert decisions and extrapolations, which is especially important for regulatory decisions in the absence of human experimental data.

Only significant features were used, which yielded fewer predictions compared to using all features, as no neighbors could be identified in many cases. However, the predictions that were actually made were reliable in most cases. We suggest that, for small datasets, the use of all features is confounding rather than informative, because there are too many insignificant features with few occurrences. Therefore, a switch to control the significance threshold for features is available in our implementation.

### 3.6 *OECD principles*

The OECD principles for (Q)SAR models [30] give conceptual advice for the evaluation of (Q)SAR methods. According to these guidelines, every model should be associated with a defined endpoint, an unambiguous algorithm, a defined domain of applicability as well as measures of robustness, goodness of fit and predictivity. If possible, a mechanistic interpretation shall be given. In our opinion, the proposed method fulfills all of these criteria:

- The algorithm has been completely described (calculation of  $p$ -values, kernel type and local prediction, see section 2), and the source code is available for the general public.
- A brief description of the modelled endpoints and datasets can be found in this article. The datasets as well as more details and background information about original data sources, and quality assurance can be found at the DSSTOX website (see References).
- The goodness of fit was assessed with the multiple correlation coefficient for all datasets, and the predictivity was assessed with two different crossvalidation techniques. The significance of the predictivity has been assessed by the response permutation procedure (see section 3.2).
- The applicability domain has been formally defined in section 2.4. It accounts for input data and response variables and is indicated by a confidence index for each prediction. The prediction quality increases nearly monotonically for all datasets with the confidence threshold, which is traded against dataset coverage. However, there is no intuitive meaning for confidence values, meaning that they differ from dataset to dataset because they depend by definition

on “density” and activity distribution of the dataset. For end users, a simple categorization (high/medium/low) is envisaged in future versions.

- Every prediction is associated with supporting information for critical evaluations and mechanistic interpretations. Due to the nature of a data-driven system it is impossible to provide mechanistic informations directly, but the presence of structurally similar compounds (neighbors) and (de)activating fragments can provide important clues. For each neighbor, the graphical user interface provides direct links to the PubChem database, where additional experimental and literature data can be retrieved. This enables an expert user to create and evaluate mechanistic hypothesis, and to check the assumption that neighbors act by similar mechanisms. In addition, activating and deactivating fragments are highlighted in red and green in the query compound for a quick identification of structural alerts. More detailed fragment visualization and query facilities are planned for future releases.

#### 4. Conclusions

This work presents a novel technique for the *in silico* prediction of toxic activities. It uses instance based support vector regression with a new activity-specific kernel, indicates the applicability domain with a confidence index for every prediction, and complies to the OECD principles for (Q)SAR models.

The whole procedure can be interpreted as an automated approach for the categorization of compounds, and the creation of local (Q)SAR models. As each prediction is based on structurally similar compounds and the presence of (de)activating substructures, it is relatively straightforward to present individual predictions in a traceable and interpretable form.

Validation studies demonstrated, that the new activity-specific kernel provides better results than the Tanimoto kernel, and that prediction accuracies increase with prediction confidences. Depending on the confidence cutoff for acceptable prediction, we were able to achieve > 90 % predictions within 1 log unit for three public datasets (Fathead Minnow Acute Toxicity, Maximum Recommended Therapeutic Dose, IRIS Lifetime Cancer Risk).

#### 5. Acknowledgements

Financial support for this work was provided by Nestec Ltd. We would like to thank P. Mazzatorta (Nestec) for discussions about the FDAMDD dataset, A.M. Richard (U.S. EPA) for providing the DSSTox datasets, and Alexandros Karatzoglou (Vienna University of Technology) for advice and discussion on the `kernlab` package [13, 28]. We appreciate the work of the free software projects gathered in the Blue Obelisk group [26].

## References

- [1] M.T. Cronin and D.J. Livingstone *Predicting Chemical Toxicity and Fate*, CRC Press, 2004.
- [2] C.G. Atkeson, A.W. Moore, and S. Schaal, *Locally Weighted Learning*, Artificial Intelligence Review 11 (1997), pp. 11–73.
- [3] C. Helma, *Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity*, Molecular Diversity (2006), pp. 147–158.
- [4] R. Guha, D. Dutta, P. C. Jurs and T. Chen, *Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions*, J. Chem. Inf. Model. 46 (2006), pp. 1836–1847.
- [5] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn and A. Tropsha, *A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models*, J. Chem. Inf. Model. 46 (2006), pp. 1984–1995.
- [6] B. Schölkopf and A.J. Smola *Learning with Kernels*, MIT Press, 2002.
- [7] S.J. Swamidass, *Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity*, Bioinformatics 21 (June 2005), pp. i359–i368(1).
- [8] J. Holliday, C. Hu, and P. Willett, *Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings.*, Comb Chem High Throughput Screen 5 (2002), pp. 155–66.
- [9] S. Kramer, L. DeRaedt, and C. Helma, *Molecular Feature Mining in HIV Data*, (2001).
- [10] C. James, D. Weininger, and J. Delany, 2000in *Daylight theory manual - Daylight 4.71* Daylight Chemical Information Systems <http://www.daylight.com>.
- [11] W.H. Press, S.A. Teukolsky, and W.T. Vetterling, 1993, 14.3. in *Numerical Recipes in C* Cambridge University Press.
- [12] J. Gower, *A general coefficient of similarity and some of its properties*, Biometrics 27 (1971), pp. 857–871.
- [13] A. Karatzoglou, A. Smola, K. Hornik and A. Zeileis, *kernelab - An S4 package for kernel methods in R*, Research Report Series / Department of Statistics and Mathematics 9 (2004), pp. 948–967.
- [14] J. Jaworska, M. Comber, C. Auer and C.J. Van Leeuwen, *Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints*, Environ Health Perspect 111 (2003), pp. 1358–1360 Congresses.
- [15] L. He and P.C. Jurs, *Assessing the reliability of a QSAR model's predictions*, Journal of Molecular Graphics and Modelling 23 (2005), pp. 503–523.
- [16] C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister and R.A. Drummond, *Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*)*, Environmental Toxicology and Chemistry 16 (1997), pp. 948–967 Dataset available online at <http://www.epa.gov/ncct/dsstox/sdf.epafhm.html>.
- [17] E.J. Matthews, N.L. Kruhlak, R.D. Benz and J.F. Contrera, *Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data*, Curr Drug Discov Technol 1 (2004), pp. 61–76 Dataset available online at <http://www.epa.gov/NCCT/dsstox/sdf.fdamdd.html>.
- [18] J.B. G.S. Backus M.A. Wolf and A. Richard, *DSSTox EPA Integrated Risk Information System (IRIS) Toxicity Review Data: SDF File and Documentation*; Dataset available online at <http://www.epa.gov/ncct/dsstox/sdf.iristr.html>.
- [19] E. Papa, F. Villa, and P. Gramatica, *Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow)*, J. Chem. Inf. Model. (2005), pp. 1256–1266.
- [20] S.P. Niculescu, A. Atkinson, G. Hammond and M. Lewis, *Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow*, SAR and QSAR in Environmental Research (2004).
- [21] T. Öberg, *A QSAR for Baseline Toxicity: Validation, Domain of Application, and Prediction*, Chem. Res. Toxicol. (2004), pp. 1630–1637.
- [22] M. Pavan, T. Netzeva, and A. Worth, *Validation of a QSAR model for acute toxicity*, SAR and QSAR in Environmental Research (2006), pp. 147–171.
- [23] R. Benigni, T. I. Netzeva, E. Benfenati, C. Bossa, R. Franke, C. Helma, E. Hulzebos, C. Marchant, A. Richard, Y.-T. Woo and C. Yang, *The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens*, J Environ Sci Health C Environ Carcinog Ecotoxicol Rev. 25 (2007), pp. 53–97.
- [24] M.J. Crawley *Statistics: An Introduction using R*, Wiley, 2005.
- [25] R. Benigni, C. Bossa, T. Netzeva and A. Worth, 2007, 4.1. in *Collection and Evaluation of (Q)SAR Models for Mutagenicity and Carcinogenicity* European Commission Joint Research Centre.
- [26] R. Guha, M. T. Howard, G. R. Hutchison, P. Murray–Rust, H. Rzepa, C. Steinbeck, J. K. Wegner and E. L. Willighagen, *The Blue Obelisk–Interoperability in Chemical Informatics*, Journal of Chemical Information and Modeling 46 (2006), pp. 991–998.
- [27] The Open Babel Package, version 2.0.1 <http://openbabel.sourceforge.net/> (accessed Feb 2006).
- [28] R-Development-Core-Team, Chapter title. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0 (2007), .
- [29] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, Robert M McDowell and P. Gramatica, *Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Perspect. 111 (2003), pp. 1361–1375.
- [30] OECD, *Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*; Available online at <http://www.oecd.org/dataoecd/33/37/37849783.pdf>.

## Tables

Table 1. LOO-CV on the EPAFHM dataset using all features. Reported are the dataset and confidence threshold, and for every combination of these criteria the number of predictions made,  $q^2$ , weighted accuracy, and mean error as well as root mean-squared error.

		SIGNIFICANCE-WEIGHTED					TANIMOTO				
Dataset	<i>conf</i>	<i>nr</i>	$q^2$	<i>wa</i>	<i>me</i>	<i>rmse</i>	<i>nr</i>	$q^2$	<i>wa</i>	<i>me</i>	<i>rmse</i>
EPAFHM	0.150	207	0.68	0.79	0.61	0.82	144	0.66	0.86	0.6	0.88
EPAFHM	0.163	195	0.68	0.78	0.62	0.83	133	0.65	0.86	0.63	0.91
EPAFHM	0.175	178	0.69	0.78	0.62	0.83	116	0.63	0.84	0.67	0.96
EPAFHM	0.188	165	0.69	0.77	0.63	0.85	100	0.68	0.87	0.62	0.88
EPAFHM	0.200	147	0.74	0.77	0.62	0.8	87	0.67	0.86	0.64	0.91
EPAFHM	0.213	127	0.75	0.76	0.65	0.83	64	0.65	0.84	0.72	1.01
EPAFHM	0.225	116	0.76	0.77	0.65	0.83	56	0.63	0.82	0.73	1.04
EPAFHM	0.238	100	0.76	0.75	0.68	0.86	45	0.62	0.84	0.72	1.07
EPAFHM	0.250	90	0.75	0.75	0.67	0.85	41	0.58	0.83	0.77	1.12
EPAFHM	0.263	81	0.72	0.74	0.68	0.87	37	0.74	0.83	0.68	0.87
EPAFHM	0.275	70	0.73	0.72	0.71	0.9	31	0.75	0.86	0.66	0.85
EPAFHM	0.288	65	0.73	0.76	0.64	0.83	27	0.68	0.87	0.62	0.8
EPAFHM	0.300	60	0.73	0.77	0.62	0.8	26	0.7	0.89	0.6	0.79
EPAFHM	median	116	0.73	0.77	0.64	0.83	56	0.66	0.86	0.66	0.91

## REFERENCES

Table 2. LOO-CV on the EPAFHM dataset, using median values across the confidence thresholds from table 1. Reported are the dataset and significance threshold, and for every combination of these criteria the number of predictions made,  $q^2$ , weighted accuracy, and mean error as well as root mean-squared error.

Dataset	sig-thr	SIGNIFICANCE-WEIGHTED					TANIMOTO				
		nr	$q^2$	wa	me	rmse	nr	$q^2$	wa	me	rmse
EPAFHM	0.00	116	0.73	0.77	0.64	0.83	56	0.66	0.86	0.66	0.91
EPAFHM	0.10	114	0.74	0.77	0.64	0.83	69	0.59	0.80	0.71	1.01
EPAFHM	0.20	118	0.73	0.76	0.65	0.84	82	0.64	0.81	0.65	0.98
EPAFHM	0.30	119	0.73	0.77	0.64	0.83	88	0.71	0.82	0.62	0.84
EPAFHM	0.40	121	0.75	0.81	0.61	0.81	104	0.72	0.82	0.58	0.80
EPAFHM	0.50	121	0.66	0.82	0.64	0.92	105	0.76	0.87	0.51	0.73
EPAFHM	0.60	105	0.78	0.88	0.51	0.73	100	0.79	0.90	0.50	0.72
EPAFHM	0.70	107	0.78	0.90	0.49	0.72	104	0.80	0.90	0.48	0.67
EPAFHM	0.80	91	0.80	0.88	0.50	0.71	88	0.79	0.87	0.52	0.74
EPAFHM	0.90	76	0.79	0.87	0.51	0.75	78	0.79	0.87	0.50	0.74

Table 3. Model comparison for Fathead Minnow acute toxicity data. Blank cells indicate "same as above". For the cited models (upper part), training ( $n_{train}$ ) and test set sizes ( $n_{test}$ ) are given together with  $q^2$  and  $rmse$  values (if reported). For this study (lower part), the data includes applicability domain estimation in the form of a confidence threshold (AD), together with test set coverage (COV). Results for the significance-weighted kernel using all features and a multilinear model based on the same neighbors are given.

REFERENCE	MODEL	$n_{train}$	$n_{test}$	$rmse$	$q^2$
Papa et al. 2005 [19]	Multil. Regr. (general model)	249	200	0.64	71%
Niculescu et al. 2004 [20]					
model 1	Neural Network	800	86	0.76	78%
<i>after excluding 20 comp.</i>		800	66	0.68	80%
model 2		800	86	1.14	52%
<i>after excluding 20 comp.</i>		800	66	0.70	80%
Öberg 2004 [21]	PLSR	208	103	-	86%
Pavan et al. 2006 [22]	Multil. Regr.	408	57	-	72%
REFERENCE	MODEL (AD, COV)	$n_{train}$	$n_{test}$	$rmse$	$q^2$
This study	Sign-w. Kernel (0.0, 100%)	459	114	0.90	53%
	Sign-w. Kernel (0.05, 90.4%)	459	103	0.84	58%
	Sign-w. Kernel (0.075, 72%)	459	82	0.66	73%
	Sign-w. Kernel (0.1, 52%)	459	59	0.68	77%
This study	Multil. Model (0.0, 100%)	459	112	0.92	54%
	Multil. Model (0.05, 90.4%)	459	100	0.87	59%
	Multil. Model (0.075, 72%)	459	60	0.73	73%
	Multil. Model (0.1, 52%)	459	64	0.70	76%

Table 4. Median values for LOO-CV and 10-CV across different confidence thresholds, using the significance-weighted kernel. Reported are the dataset,  $r^2$ ,  $nr$ ,  $q^2$  (only for LOO-CV),  $wa$ ,  $me$  and  $rmse$ . For every dataset, the upper row shows the results for all features, the lower for significant features only.

			LOO-CV					10-CV			
Dataset	Ft.	$r^2$	$nr$	$q^2$	$wa$	$me$	$rmse$	$nr$	$wa$	$me$	$rmse$
EPAFHM	all	0.73	165	0.69	0.77	0.64	0.85	14	0.8	0.69	0.89
EPAFHM	sign.	0.78	76	0.79	0.87	0.51	0.75	12	0.85	0.61	0.8
IRIS	all	0.5	22	0.37	0.81	0.78	1.07	-	-	-	-
IRIS	sign.	0.66	17	0.77	0.96	0.41	0.56	-	-	-	-
FDAMDD	all	0.70	319	0.67	0.92	0.43	0.59	27.5	0.9	0.48	0.61
FDAMDD	sign.	0.69	313	0.7	0.92	0.43	0.6	29	0.94	0.47	0.61

## Figures

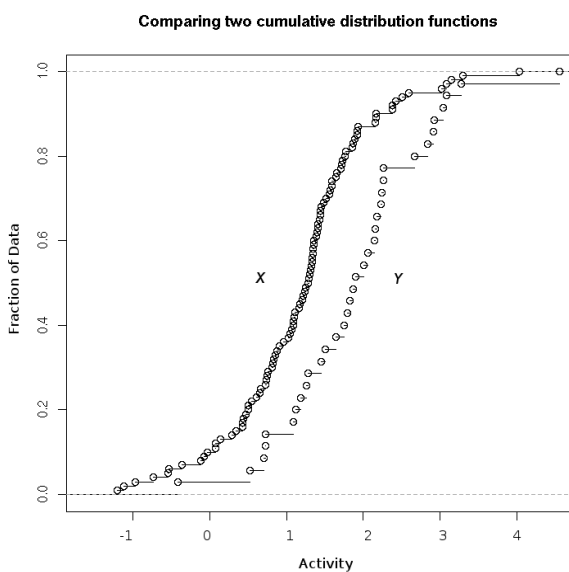


Figure 1. Comparison of the cumulative activity distributions of two (hypothetic) sets of activity values  $X$  and  $Y$  with sizes 100 and 35, respectively. The mean value of  $X$  is 1.0, the mean value of  $Y$  is 2.0. It is highly unlikely ( $p = 0.0001319$ ) that  $X$  and  $Y$  have been drawn from the same data source. If, for a feature  $f$ ,  $Y = A_{|f}$  and  $X = A$ ,  $f$  would be significantly deactivating with a  $p_f = 1 - 0.0001319 = 0.9998681$ .

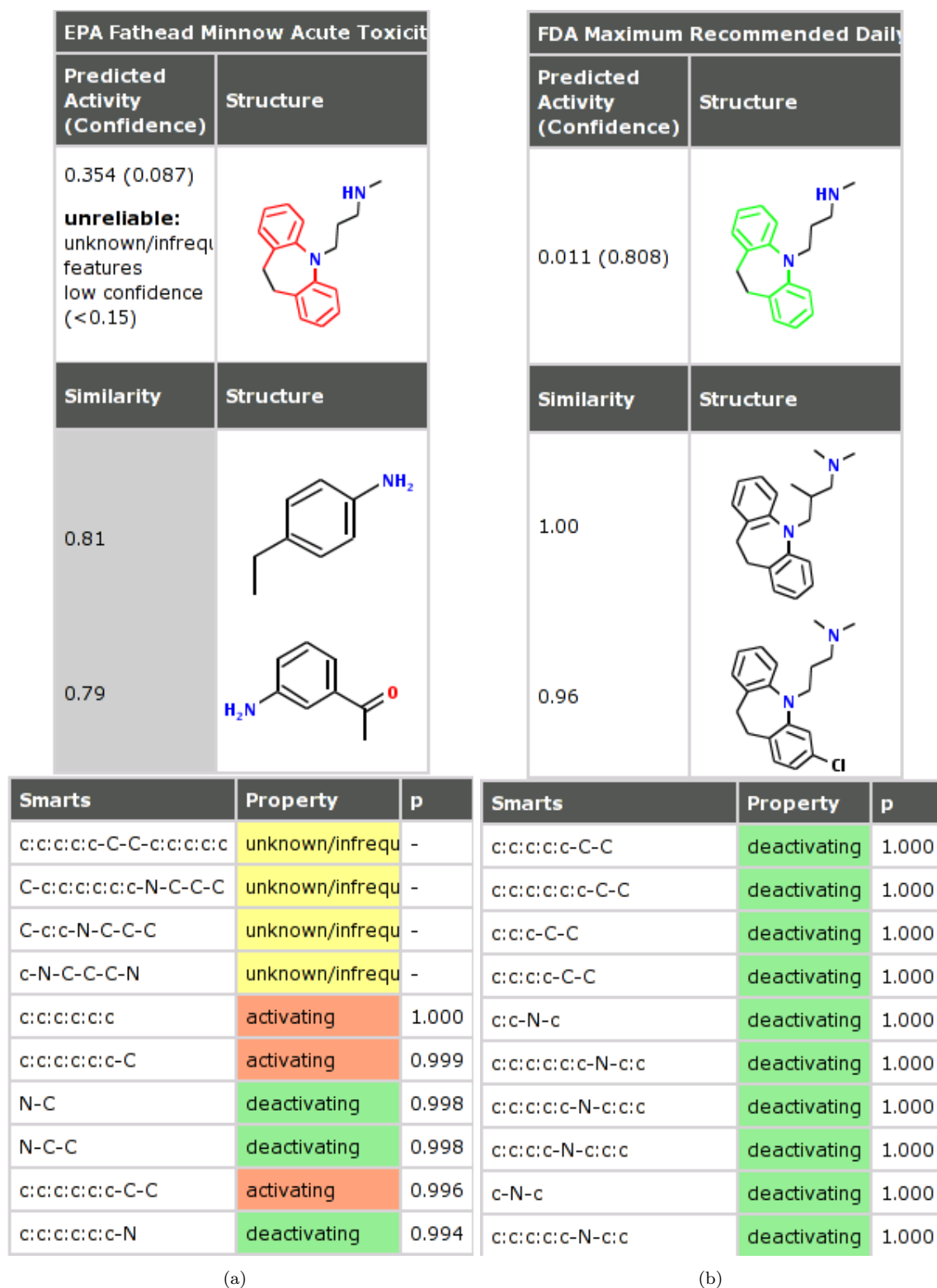


Figure 2. Query structure CNCCCNC1CCCC1CCc2cccc23 (Desipramine, top), and the two most similar neighbors as well as the 10 most significant fragments for the EPAFHM (a) and FDAMDD (b) dataset. The similarity is determined by *relevant* fragments shared with the query structure. Those are different for the two endpoints. The fragments are additionally color-coded: red and green for activating and deactivating, and yellow for unknown fragments.

REFERENCES

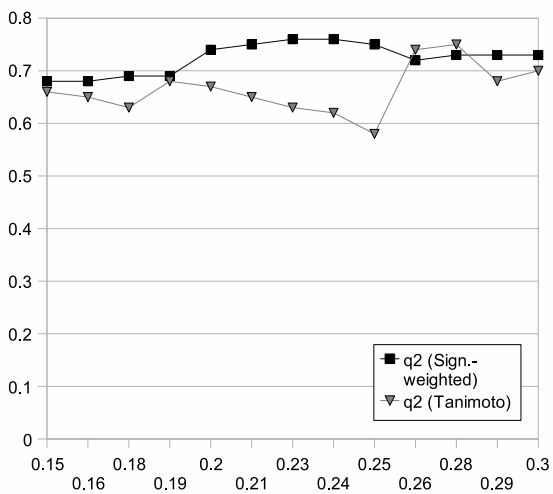


Figure 3. Comparison of the multiple correlation coefficient ( $q^2$ ), obtained by LOO-CV for the EPAFHM dataset, plotted over different confidence thresholds (see Table 1).

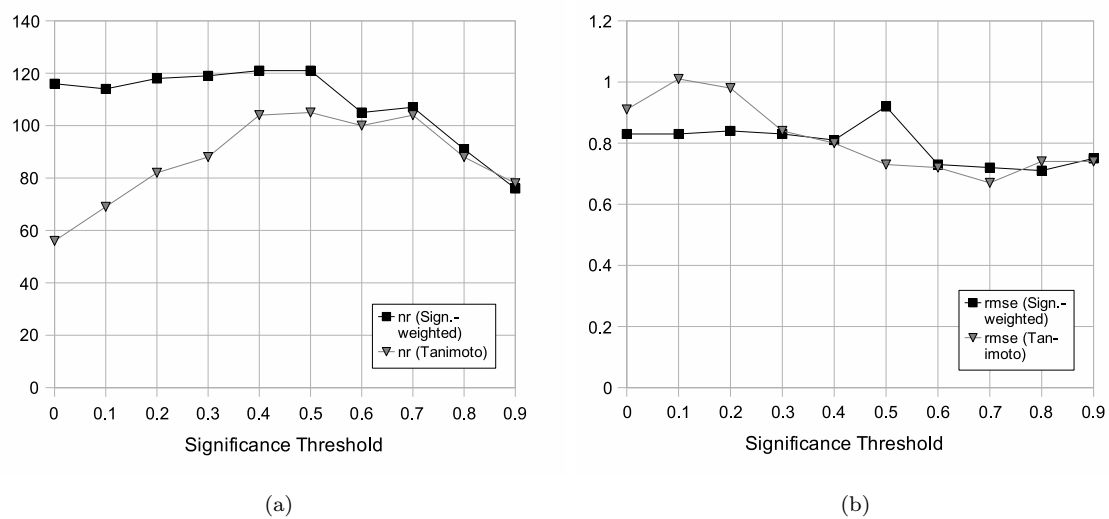


Figure 4. Comparison of the median  $nr$  (a) and  $rmse$  (b) values, obtained by LOO-CV for the EPAFHM dataset, plotted over different significance thresholds (see Table 2).

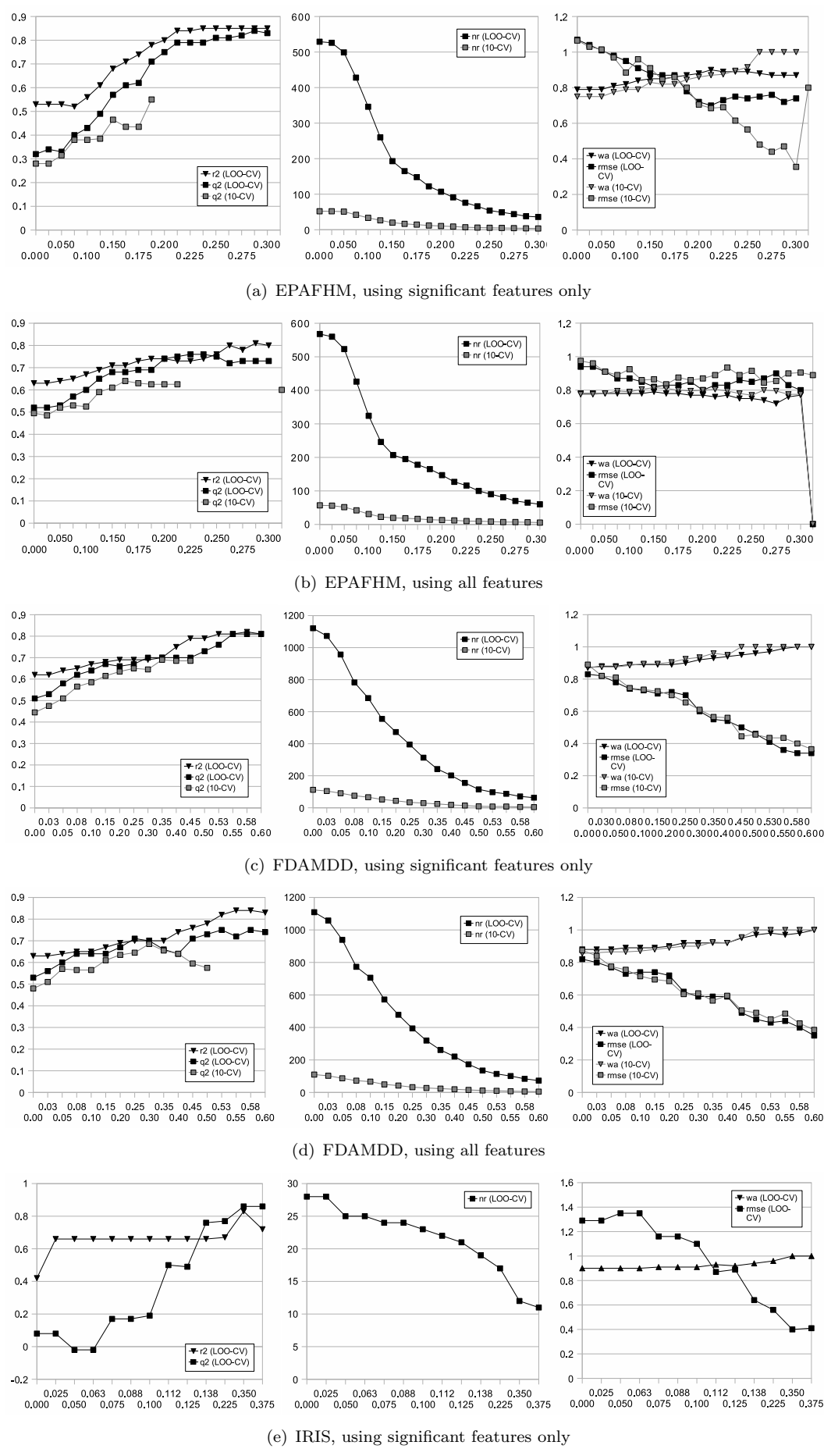


Figure 5. Predictive validation plots for the EPAFHM (a, b), FDAMDD (c, d), and IRIS (e) datasets, using significant features only (a, c, e) and all features (b, d). Reported are  $r^2$ ,  $q^2$ ,  $nr$ ,  $rmse$  and  $wa$ .

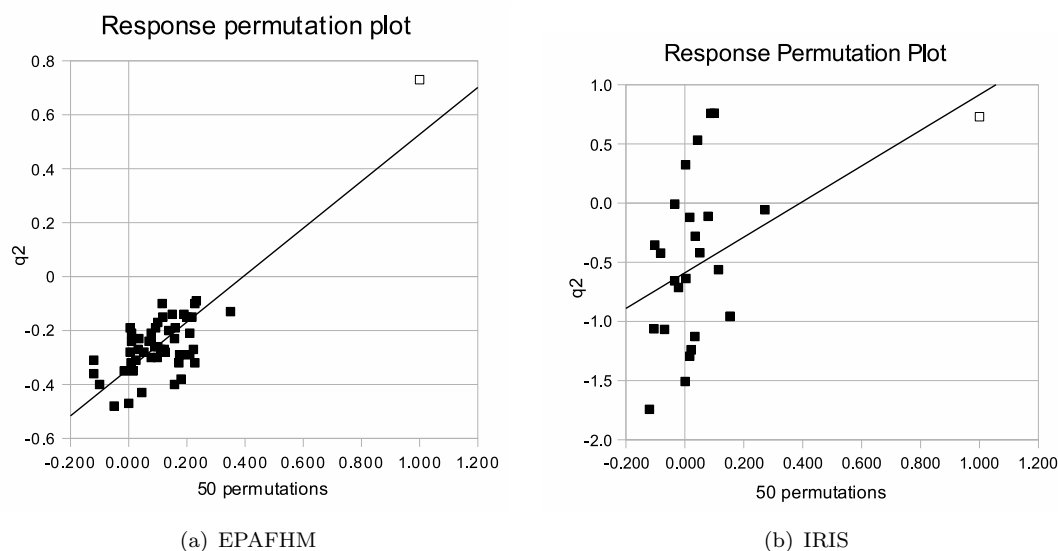


Figure 6. Response permutation testing for the EPAFHM (a) and IRIS datasets (b):  $q^2$  values obtained by using permuted (filled dots) and original activities (outlined dot), plotted over activity correlation with the original values. For the IRIS dataset, some points have the same coordinates. Due to the small dataset, the values for the IRIS dataset vary highly. However, the intercept is clearly below 0.05 in both cases, indicating a valid model [29].