RESEARCH ARTICLE

Open Access

Defining a novel *k*-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions

Faizan Sahigara, Davide Ballabio, Roberto Todeschini and Viviana Consonni[*]

Abstract

Background: With the growing popularity of using QSAR predictions towards regulatory purposes, such predictive models are now required to be strictly validated, an essential feature of which is to have the model's Applicability Domain (AD) defined clearly. Although in recent years several different approaches have been proposed to address this goal, no optimal approach to define the model's AD has yet been recognized.

Results: This study proposes a novel descriptor-based AD method which accounts for the data distribution and exploits *k*-Nearest Neighbours (kNN) principle to derive a heuristic decision rule. The proposed method is a three-stage procedure to address several key aspects relevant in judging the reliability of QSAR predictions. Inspired from the adaptive kernel method for probability density function estimation, the first stage of the approach defines a pattern of thresholds corresponding to the various training samples and these thresholds are later used to derive the decision rule. Criterion deciding if a given test sample will be retained within the AD is defined in the second stage of the approach. Finally, the last stage tries reflecting upon the reliability in derived results taking model statistics and prediction error into account.

Conclusions: The proposed approach addressed a novel strategy that integrated the kNN principle to define the AD of QSAR models. Relevant features that characterize the proposed AD approach include: a) adaptability to local density of samples, useful when the underlying multivariate distribution is asymmetric, with wide regions of low data density; b) unlike several kernel density estimators (KDE), effectiveness also in high-dimensional spaces; c) low sensitivity to the smoothing parameter *k*; and d) versatility to implement various distances measures. The results derived on a case study provided a clear understanding of how the approach works and defines the model's AD for reliable predictions.

Keywords: QSAR, Applicability domain, kNN, Nearest neighbour, Model validation

Background

The popularity of QSARs has seen a growth from time to time and was complemented by the availability of more sophisticated and efficient model development techniques. This fact was further supported by the consideration of QSAR predictions for regulatory purposes. To deal with risk assessment of chemicals for their safe use, a new European legislation – REACH (Registration, Evaluation, Authorization and restriction of Chemicals) was approved in the recent years [1]. To reduce animal testing and replacing them by cost effective methods, this law encourages the use of QSARs as a possible alternative when enough experimental data is not available, provided that the model was strictly validated for its regulatory consideration [2].

There are several aspects that must be taken into account before considering a QSAR model reliable enough. In other words, the validity of a model has to be evaluated. Existing literature has often emphasized upon validating the QSAR models to reflect their robustness and predictive ability. In 2004, following five OECD principles for model validation were adopted to validate a QSAR model for its regulatory consideration: a) a defined endpoint; b) an unambiguous algorithm; c) a defined domain of applicability d) appropriate measures for goodness-of-fit, robustness and predictivity and e) mechanistic interpretation, if possible [3].

* Correspondence: viviana.consonni@unimib.it
Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, Milano 20126, Italy

Applicability domain (AD) of a QSAR model defines the model's limitation in its structural domain and response space. In other words, this principle for model validation restricts the applicability of a model to reliably predict those test samples that are structurally similar to the training samples used to build that model [4-6]. Several approaches were proposed in the past years to define the AD of QSAR models. These approaches mainly differed in the algorithm used to characterise the AD within the descriptor space, where the model can predict reliably [7,8]. For instance, some classical approaches suggested defining the domain of applicability by a) considering the range of descriptors values; b) enclosing the training space in a convex hull; c) calculating the distance of a query compound from a defined point within the model's descriptor space and d) estimating the Probability Density Function for the given data. All these approaches were associated with their own advantages and limitations [2,7-10]. From time to time, several approaches were proposed that were aimed to be more efficient or were thought to overcome several limitations of existing approaches.

This article proposes a new heuristic approach towards defining the AD of QSAR models. The basis of this novel strategy is inspired from the *k*-Nearest Neighbours (kNN) approach and adaptive kernel methods for probability density estimation (kernel density estimators, KDE) [11]. Due to its simplicity and easy implementation, kNN had been a preferred choice for several proposed QSAR studies [6,12-18].

In the classical kNN approach for AD evaluation [6,18], average distances of all the training samples from their *k* nearest neighbours are calculated and used to define a unique threshold to decide if a test sample is inside or outside the model's AD (for example, 95th percentile). Moreover, in the framework of the probability density function estimation, the nearest neighbour method provides density estimates depending on the Euclidean distance to the *k*-th nearest data point [19]. Following the same concept, the proposed method tries to integrate the kNN principle with the salient features of adaptive kernel methods [11], which define local bandwidth factors corresponding to the training data points and use them to build the density estimate at a given point.

The novelty of the kNN based AD approach proposed in this article lies in the overall strategy that is properly executed in a three-stage procedure to encapsulate and reflect upon several significant aspects towards model validation. Moreover, some features common to most of the AD approaches were dealt differently with this approach; for instance, rather than defining a general threshold as in all the distance-based approaches, each training sample in this approach was associated with its individual threshold; in order to find an optimal smoothing parameter *k*, this

approach performed a *k*-optimization procedure based on Monte Carlo validation; additionally, model's statistical parameters and other relevant aspects were dealt simultaneously to reflect upon the reliability in the derived results.

To better understand the strategy behind this approach, it was implemented on a dataset from the literature. The dataset was chosen from the CAESAR project to predict the bioconcentration factor (BCF) [20,21].The derived results were discussed in comparison with those derived from other literature AD approaches.

Methods
k-Nearest Neighbours principle from AD perspective
The kNN principle basically reflects upon the structural similarity of a test sample to the training samples used to build that model. In theory, the distance of a query sample is considered from its *k* closest data points in the chemical space. Lower distance values correspond to a higher similarity, while the increasing distances signify higher levels of structural mismatch. The *k* value plays a significant role in defining how constraint the approach will be and thus, it can be referred to as the smoothing parameter.

A stepwise execution of the following three stages characterises the workflow of this approach:

1) defining thresholds for training samples
2) evaluating AD for new/test samples
3) optimizing the smoothing parameter *k*

To allow a better interpretation of the proposed approach, results on a two-dimensional simulated dataset will be considered throughout the major part of this discussion and wherever applicable. As shown in Figure 1, this dataset has a cluster of 48 training samples and the remaining two training samples (49 and 50) are located quite in the extremities of the space with respect to these clustered samples.

Defining thresholds for training samples
Thresholds have a great influence in characterising the AD for reliable predictions; a test sample that exceeds the threshold condition is associated with an unreliable prediction.

Like the adaptive kernel methods, instead of defining a general unique threshold as seen with several classical AD approaches, the proposed approach allocates a set of thresholds corresponding to the various training samples.

For a given value of *k*, threshold allocation process can be summarised as follows:

a) First of all, the distances of each training sample from the remaining *n* − 1 samples are calculated and ranked in increasing order, *n* being the total number of training samples. This will result in a *n* x (*n* −1)

Figure 1 Scatter plot of the simulated dataset.
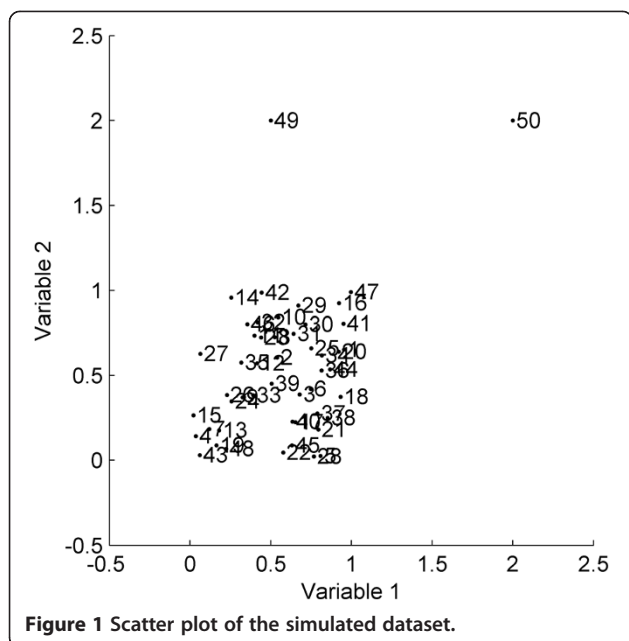
neighbour table **D**; an entry D_{ij} of the table corresponds to the distance of the i-th sample from its j-th nearest neighbour:

$$D_{i1} \le D_{i2} \le \dots \le D_{i,n-1}$$

b) The average distance of each i-th sample from its k nearest neighbours is calculated considering the first k entries in i-th row of the neighbour table:

$$\bar{d}_i(k) = \frac{\sum_{j=1}^{k} D_{ij}}{k}$$
$$where, 1 \le k \le n-1 \ \text{and} \ \bar{d}_i(k) \le \bar{d}_i(k+1) \tag{1}$$

A vector $\bar{\mathbf{d}}(k)$ of average distance values is then derived considering all the samples in the training set.

c) Next, a reference value (from now on referred as *Ref Val*), $\tilde{d}(k)$ is determined as follows:

$$\tilde{d}(k) = Q3(\bar{\mathbf{d}}(k)) + 1.5[Q3(\bar{\mathbf{d}}(k)) - Q1(\bar{\mathbf{d}}(k))] \tag{2}$$

where, $Q1(\bar{\mathbf{d}}(k))$ and $Q3(\bar{\mathbf{d}}(k))$ are the values corresponding to the 25th and 75th percentiles in the vector $\bar{\mathbf{d}}(k)$, respectively [22].

d) Next, the ordered distances of each i-th training sample from all other n - 1 training samples are compared with the *Ref Val*. If the distance value of the i-th sample from its given j-th training neighbour (where $1 \le j \le n-1$) is less than or equal to the *Ref Val*, then that distance value is retained, otherwise is discarded. The number K_i of neighbours

satisfying this condition, minimum zero and maximum being $n - 1$, defines the density of the i-th sample neighbourhood:

$$K_i: \quad \{D_{ij} \le \tilde{d}(k)\} \quad \forall j : 1, n-1 \tag{3}$$

e) Finally, each i-th training sample is associated with a threshold t_i which defines the width of its neighbourhood as:

$$t_i = \frac{\sum_{j=1}^{K_i} D_{ij}}{K_i} \tag{4}$$

If no distance value was retained for a given i-th training sample ($K_i = 0$), then its threshold t_i would be theoretically settled to 0, but a pragmatic solution is to set it equal to the smallest threshold of the training set.

The plot in Figure 2 provides with an overview of the thresholds for all the 50 samples in the simulated dataset. As expected, most of the training samples within the cluster (for instance, samples 2, 33 and 39) were associated with higher K_i values. On the other hand, obvious potential outliers (samples 49 and 50) had their thresholds equal to 0 since they couldn't satisfy the threshold criterion even for a single training neighbour (i.e. $K_i = 0$), thus no distance values contributed to their threshold calculation. Nevertheless, they were associated with the minimum threshold equal to 0.42, i.e. the threshold of sample 43.

Figure 2 Simulated data set. Thresholds t_i vs. number of training neighbours K_i plot ($k = 12$).
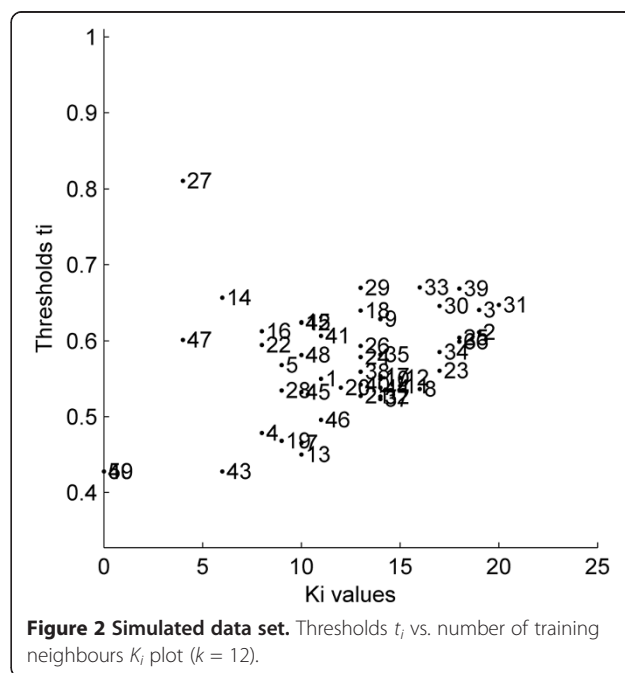
Evaluating AD for new/test samples

Until this point, each training sample was associated with its individual threshold. The next step will be to characterise the AD which usually relies upon a set of conditions that will decide if a given test sample can be associated with a reliable prediction or not.

The criterion used by this approach to associate a given test sample to be within the domain of applicability can be summarised below.

Given a test sample, its distance from all the n training samples is calculated and simultaneously, compared to be less than or equal to the thresholds associated with those training samples. If this condition holds true with at least one training sample, the test sample will be considered inside the domain of applicability for that model. Otherwise, the prediction for that test sample will be rendered unreliable.

More formally, given the training set TR, for each test sample j, the AD decision rule is:

$$j \in AD \quad iff \quad \exists i \in TR : \quad D_{ij} \le t_i \quad (5)$$

where D_{ij} is the distance between the j-th test sample and the i-th training sample and t_i is the individual threshold of the latter. In addition, each test/new sample will be associated with the number K_j of nearest training neighbours for which the previous condition holds true. This number can be assumed as a measure of prediction reliability; indeed, high values of K_j indicate that the new sample falls within a dense training region of the model's space, while low values of K_j denote that the new sample still belongs to the model's space, but located in sparse training regions. K_j equal to zero rejects the sample as it being outside the model's AD since no training neighbours are identified.

Figure 3 provides with the contour plot for the simulated dataset derived projecting several data points enough to fill the training space. Thresholds were calculated using 12 nearest neighbours and Euclidean distance. This choice of k = 12 nearest neighbours was based on the results derived performing an internal k-optimization, discussed later in this article. The space enclosed around the cluster represented as black line indicates that all the data points within this enclosed region were inside the AD. Thus, this region reflects in a way how the AD was characterised for this two-dimensional dataset. Area of this enclosed region tends to expand or shrink depending upon the number of nearest neighbours used for threshold calculation.

As explained earlier, the extreme outliers in the training space will be associated with the number K_i of neighbours equal to zero and the lowest possible threshold in the training set. Consider the sample 49 from the simulated dataset which is an extreme outlier with its threshold equal to 0.42. If there is a test sample that seems to
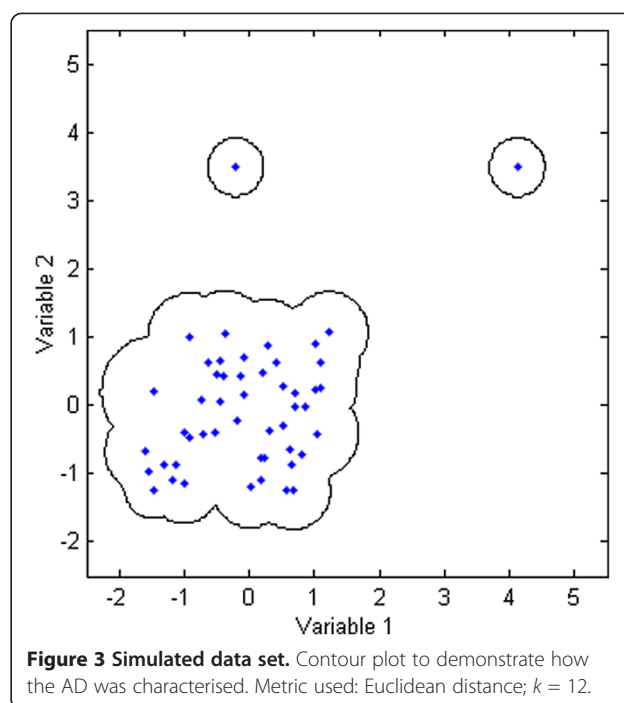

Figure 3 Simulated data set. Contour plot to demonstrate how the AD was characterised. Metric used: Euclidean distance; k = 12.

be quite in the vicinity of this potential outlier within the descriptor space, the test sample will be associated with an unreliable prediction since its distance from sample 49 will likely exceed the small threshold. Now, consider a case, where the descriptor values for another test sample exactly overlap or are very similar to those for this potential outlier. In this situation, the distance of that sample from the outlier will be less than the threshold and thus it will be considered within the domain of applicability. In theory, this is not wrong because the potential outlier is still a part of the training space. Practically, the approach retains all the training samples to characterize the AD but minimizing the role of potential outliers in doing so. That's the reason why the first test sample was excluded from being reliably predicted while the second sample was not. However, for the latter the number K_j of nearest training neighbours will likely be equal to one indicating that its prediction has some degree of uncertainty. In conclusion, there exists a relation between the defined AD and the impact of training samples in characterising it based on their threshold values.

Optimizing the smoothing parameter k

Another important aspect is concerning the choice of an appropriate smoothing parameter k, whose theoretical range is between 1 and n-1.

Very low k values will restrict the domain of applicability in a very strict manner as compared to the AD derived opting for larger k values. This is because, an opted k value will have a direct impact on the threshold calculations which in turn can make it more rigid or easier

for test samples to satisfy the threshold criterion. The strategy implemented in this article to select an appropriate k value was performed by Monte Carlo validation, maximizing the percentage of the test samples considered within the AD, i.e. satisfying AD criterion (Equation 5).

Box-and-whisker plots (box plots) were produced to get an overview of all these derived results. For instance, consider the plot in Figure 4 derived for the simulated dataset showing percentage of test samples retained within the AD with different k values (optimization carried out with 20% of samples in the test set and 1000 iterations).

Median quartile in the middle of the box (marked in red) can be referred for all the k values to get a hint about how many test samples were retained on average during the optimization process for a given k value. The top and bottom edges of each box plot (quartiles $Q3$ and $Q1$) correspond to 75th and 25th percentile, respectively. The whisker can extend further from $Q1-w(Q3-Q1)$ until $Q3+w(Q3-Q1)$, of 1.5 [23]. The test samples falling outside this coverage are considered as outliers and are highlighted as '+' in red. About their usefulness in the proposed AD approach, box plots showing limited spread and allowing majority of test samples to be retained within the AD can be favoured and their corresponding range of k values can be considered to finally opt for the most appropriate k. Additionally, a line plot is integrated in the same figure indicating the mean percentage of test samples that were considered within the AD for each k value. A simultaneous interpretation of both these plots can make it easier for a user to decide upon an appropriate k value.

Figure 4 shows that the spread of the box plots for initial k values is quite large. This may have resulted due to
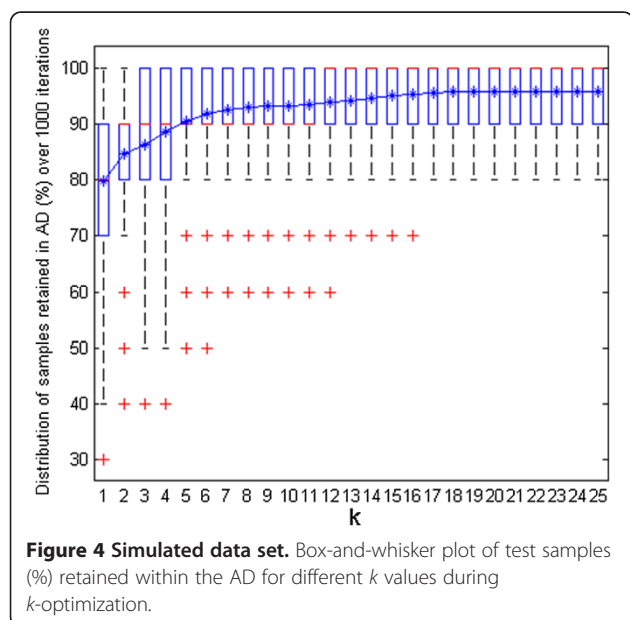
the impact of restricted training thresholds that excluded several test samples from the AD. With an increase in k values, the spread narrowed, however the outliers were still present until $k = 17$. After this point, the box plots remained unchanged throughout the plot with no outliers. Similar observations were derived from the mean line plot which showed a significant rise initially followed by a stable curve until the first half of the k values. The plot didn't show any major changes for the second half of the k values. In order to avoid very high k values good enough to unnecessarily expand the defined AD, a k value of 12 was opted as appropriate k for this dataset. The plots dealt earlier (Figures 2 and 3) for this dataset were thus derived using this opted k value.

We also performed an extended analysis on several diverse data sets (results not reported in this paper), to study the influence of the smoothing parameter k on model's AD definition. It was concluded that optimization of k can be a time-demanding procedure especially in the case of a huge number of samples, but it was also observed that this approach is quite insensitive to the smoothing parameter k, except for very small k values which led to the results influenced by local noise. Therefore, for many applications the optimization of the smoothing parameter can be avoided and reasonable results can instead be obtained by a fixed k value empirically calculated as $n^{1/3}$.

Reflecting the reliability in derived results

After the AD approach has been applied to the model of interest, several features will be taken into account to reflect upon the derived results. Moreover, as stated earlier the response domain will be taken into account to address the reliability in the results derived by characterising the AD of a model in its descriptor space.

In order to reflect upon a model's predictive ability, the predictive squared correlation coefficient (Q^2) was used. Since the test samples excluded from the model's AD are unreliably predicted, in theory they should not be accounted for to calculate the model's statistics (Q^2).

The following key parameters were evaluated:

a) Number of test samples retained within the AD.
b) Q^2 calculated from the test samples retained within the AD [24,25]:

$$Q^2 = 1 - \frac{\left[\sum_{j=1}^{n_{TS}}\left(\hat{y}_j - y_j\right)^2\right]/n_{TS}}{\left[\sum_{i=1}^{n_{TR}}\left(y_i - \bar{y}_{TR}\right)^2\right]/n_{TR}} \qquad (6)$$

where, y_j is the measured response value for the j-th sample and $\hat{y}_j$ its predicted value; n_{TR} and n_{TS} represent the total number of training and test

Figure 4 Simulated data set. Box-and-whisker plot of test samples (%) retained within the AD for different k values during k-optimization.

samples, respectively, and $\bar{y}_{TR}$ is the mean response of the training set.

c) List of all the test samples considered outside the AD.

d) For each j-th test sample, the absolute standardized error calculated as:

$$SE_j = \frac{\left|y_j - \hat{y}_j\right|}{s_Y} \qquad (7)$$

where, y_j is the measured value for the j-th sample and $\hat{y}_j$ its predicted value; s_Y the standard error of estimate derived from the training set.

e) The information about how many times the threshold criterion (Equation 5) is satisfied by each test sample, that is, how many training neighbours (i.e. K_j) are located at a distance less than or equal to their threshold values, from a given test sample.

In theory, a test sample satisfying the threshold criterion several times (i.e. having high K_j) is expected to be predicted with higher accuracy. This can be desired since less distant training neighbours indicate a higher structural similarity of the test sample. On the contrary, a test sample satisfying the threshold criterion for no training neighbours ($K_j = 0$) indicates that there wasn't any training sample similar enough to reliably predict that test sample.

Results and discussion

As the case study to derive results with the proposed strategy, the CAESAR Model 2 to predict bioconcentration factor (BCF), which was developed under the EU project CAESAR following the REACH requirements, was selected. It is a Radial Basis Function Neural Network (RBFNN) model derived from 378 training and 95 test samples [20,21]. The five descriptors used to develop this model were calculated using Dragon 5.5 [26].

The statistics for this model are summarized in Table 1.

For comparison purposes, some AD approaches taken from literature [2,7-10] were implemented on the selected case study. Among them, the classical kNN-based AD approach [6,18] was implemented by calculating average distances of all the training samples from their 5 nearest neighbours (i.e. $k = 5$); since the choice of

thresholds didn't follow any strict rules in the existing literature, the value corresponding to 95th percentile in this vector of average distances was considered as general threshold. If the average distance of a test sample was lesser than or equal to the threshold value, the test sample was retained within the AD.

In addition to the classical kNN-based AD approach, the following methodologies were considered [2,7-10]: the Bounding Box, which is based on the ranges of model variables; its variant based on principal components instead of the original variables (PCA-Bounding Box); the Convex Hull, which is the smallest convex area that contains the original set; two distance-based methods, which calculate the distance (Euclidean and Mahalanobis) of a test sample from the data centroid and use the 95th percentile of the training sample distances as threshold.

Finally, some methods for probability density function estimation were also considered. Among the multivariate kernel density methods, four variants of Gaussian kernel estimators were implemented [19]: fixed Gaussian kernel with bandwidth equal to 0.462 (for the studied data set); optimized Gaussian kernel with a smoothing parameter equal to 0.237 obtained by leave-one-out cross-validation [27]; variable Gaussian kernel with bandwidth calculated as the inverse function of the Euclidean distance to k-th neighbour ($k = 15$) [27]; adaptive Gaussian kernel, with fixed Gaussian kernel as the pilot estimate and sensitivity parameter α equal to 0.5. Finally, Epanechnikov kernel with a fixed bandwidth equal to 1.961 and the nearest neighbour density estimator with smoothing parameter k equal to 15, were also considered [19].

For all the implemented methods, except for Bounding Box and Convex Hull, autoscaling was adopted as data pretreatment.

The proposed AD strategy was implemented in MATLAB [28] using autoscaled Euclidean distances. The k-optimization procedure was carried out initially to decide upon an optimal k value; the training set of 378 samples was randomly partitioned 1000 times selecting 20% of samples in the test set (i.e. 75 samples). The box plots in Figure 5 summarize the percentage of test samples retained within the AD for different k values (up to 25).

As expected, the first lower k values were associated with box plots having highest spread. This degree of dispersion lowered gradually with increase in number of neighbours considered. The line plot of the mean showed an increase in the number of samples throughout the plot, however, this increment after initial k values was gradual. Based on their lower spread and preference to retain reasonably higher number of samples within the AD (as reflected from their median), the k values in the range of 15–19 were considered further to decide upon an optimal k. Finally, to avoid unnecessarily higher training thresholds and their resulting impact on the defined AD, $k = 15$ was

Table 1 Summary of model statistics for the case study

Model	Training set			Test set		
	n_{TR}	R^2	RMSE	n_{TS}	Q^2	RMSEP
CAESAR Model 2	378	0.804	0.591	95	0.797	0.600

R^2 Determination coefficient; *RMSE* Root-mean-square error; Q^2 Predictive squared correlation coefficient; and *RMSEP* Root-mean-square error of prediction.
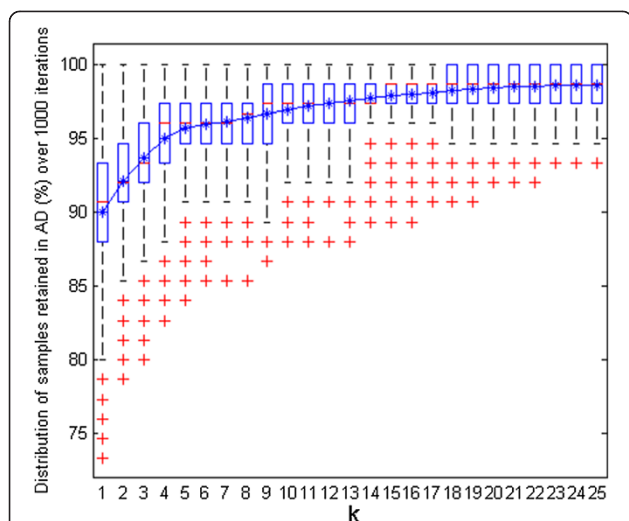
Figure 5 CAESAR BCF model. Box-and-whisker plot of test samples (%) retained within the AD for different k values during k-optimization.

considered as an optimal choice for this case study. Considering the selected k value of 15, the novel AD approach identified four test samples (33, 61, 82 and 83) being outside the model's AD.

To reflect upon the reliability in the results derived with this approach, absolute standardized error of all the test samples was plotted against their corresponding K_j values. As shown in Figure 6, four test samples considered outside the AD with this approach were associated with a value of $K_j = 0$. The absolute standardized error for sample 33 was quite higher as compared to the remaining three samples. As seen clearly from the plot,
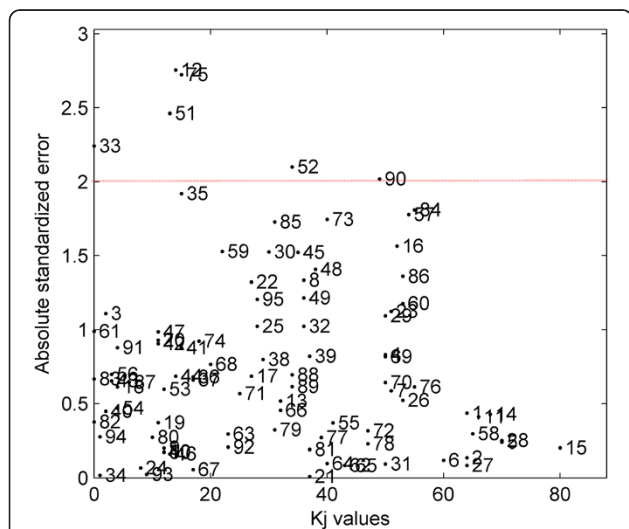
Figure 6 CAESAR BCF model. Absolute standardized error of test samples plotted against their K_j values.

there had been a sharp decrease in the prediction error of test samples with an increase in K_j. However, it can't be denied that this pattern wasn't rigidly followed in the results. There had been test samples with very low K_j values but extremely low or negligible absolute standardized error, meaning that even less reliable predictions can have good accuracy. In any case, this plot somehow tried to interpret the AD results derived in model's descriptor space taking into account the response domain, and clearly informed about both, reliability and accuracy in the predictions of the test samples. Most of the predictions had good accuracy, them being within two units of standardized error and high reliability. The samples corresponding to these reliable predictions were associated with higher K_j values, thus being well represented by several training samples.

Generally, a standardized error of two/three units is usually considered as a warning value for outliers detection. In Figure 6, six test samples (12, 33, 51, 52, 75 and 90) exceeded a two-unit threshold for the absolute standardized error indicating them as outliers in the model's response domain. It can be interesting to further evaluate the reasons behind categorising them as outliers; however, this is beyond the scope of this article as the proposed AD approach is defined within the model's descriptor space. Nevertheless, this evaluation identifies sample 33 as an outlier in model's descriptor's space as well as its response domain which further supports the results derived from the proposed approach to exclude this sample from the model's AD.

Finally, the results derived by this approach were compared with those derived from classical AD approaches. Table 2 reports these results; the first row shows results when no AD approach has been applied to bound the model's descriptor space.

The number k of nearest neighbours considered with the proposed approach (i.e. 15) was comparatively higher than the one considered with classical kNN (i.e. 5); however, the impact on model statistics was not so obvious on the resulting Q^2, while the number of retained samples increased from 87 (classical kNN) to 91 (proposed approach). Discussing the results derived with classical approaches, number of samples retained within the AD varied significantly depending on what strategy was used. Convex hull, optimized and variable Gaussian kernel methods retained the least number of samples while the Bounding Box considered none of the test samples outside the AD. Overall, the proposed approach worked quite well on the CAESAR model, trying to define an AD with maximum retained test samples within the domain and positive impact on the model statistics.

The last column of Table 2 reports the list of samples considered outside the AD with all the approaches. Irrespective of total number of samples considered outside

Table 2 Comparison of AD methods applied to the test set of CAESAR BCF model

Approach	IN AD	Q^2	OUTSIDE AD
All samples inside (no AD approach)	95	0.797	None
Proposed approach (Euclidean dist., $k = 15$)	91	0.803	33 61 82 83
Bounding box	95	0.797	None
PCA bounding box	93	0.804	33 40
Convex hull	73	0.789	3 7 9 13 18 33 34 36 37 38 39 40 41 43 51 56 61 72 79 91 92 94
Euclidean dist (95 percentile)	88	0.802	3 33 36 37 40 42 61
Mahalanobis dist (95 percentile)	89	0.791	18 43 54 61 83 91
Classical kNN (Euclidean dist., $k = 5$)	87	0.797	3 33 34 40 61 82 83 94
Fixed Gaussian kernel	85	0.794	3 24 33 34 40 61 82 83 91 94
Optimized Gaussian kernel	66	0.831	3 912 22 24 33 34 38 40 45 47 51 53 54 56 61 68 69 75 76 80 82 83 87 89 91 93 94 95
Variable Gaussian kernel ($k = 15$)	81	0.790	3 24 33 34 40 43 61 80 82 83 89 91 94 95
Adaptive Gaussian kernel	88	0.801	3 33 43 61 82 83 91
Fixed Epanechnikov kernel	87	0.799	3 33 40 43 61 83 91 94
Nearest neighbour density estimator ($k = 15$)	91	0.806	3 33 61 91

the AD, all the methods converged significantly identifying a subset of common samples that were always excluded from the model's AD.

Conclusions

A novel kNN-based approach to define the AD of QSAR models was proposed. The overall execution of this approach was performed in three different phases that efficiently used the salient features of kNN principle to define a model's AD in its descriptor space. Significant features that distinguished the proposed AD approach include defining individual threshold for each training sample, optimizing the smoothing parameter k to be considered and taking into account the model's response domain to reflect upon the reliability of results derived in its descriptor space.

In the proposed AD method, the appropriate number k of neighbours can be chosen on the basis of the plot with retained samples vs. k values obtained by Monte Carlo validation; it allowed to identify a smoothed region of the k values where the results remained unchanged, ensuring high robustness in the AD definition.

The results on the selected case study defined an AD with a positive impact on model statistics retaining maximum possible samples that were reliably predicted. Comparison of the derived results with those from the classical approaches by no means intended to project the pitfalls of existing approaches but it was aimed to have a performance evaluation of this novel strategy to understand how its implementation could lead to obtain similar or different results as compared to the classical

ways of defining the AD. An extended comparison of the different AD approaches on several diverse data sets have indicated the following relevant features that characterize the proposed AD approach: a) adaptability to local density of samples, useful when the underlying multivariate distribution is asymmetric, with wide regions of low data density; b) unlike several kernel density estimators, effectiveness also in high-dimensional spaces; d) low sensitivity to the smothing parameter k; d) versatility to implement various distances measures other than Euclidean distance, such as Manhattan distance, Mahalanobis distance and the recently proposed locally-centred Mahalanobis distance [29], depending on the data set in analysis.

A MATLAB module for the model's AD estimate by different approaches will be soon available at http://michem.disat.unimib.it/chm/.

References

1. *REACH. European Community Regulation on chemicals and their safe use.* http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm.
2. Worth AP, Bassan A, Gallegos A, Netzeva TI, Patlewicz G, Pavan M, Tsakovska I, Vracko M: *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance. ECB Report EUR 21866 EN, 95pp.* Ispra, Italy: European Commission, Joint Research Centre; 2005.
3. *OECD. Quantitative Structure-Activity Relationships Project.* http://www.oecd.org/document/23/0,3746,en_2649_34377_33957015_1_1_1_1,00.html.
4. Worth AP, van Leeuwen CJ, Hartung T: **The prospects for using (Q)SARs in a changing political environment: high expectations and a key role for the Commission's Joint Research Centre.** *SAR QSAR Environ Res* 2004, **15**:331–343.
5. Nikolova-Jeliazkova N, Jaworska J: **An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN.** *Altern Lab Anim* 2005, **33**:461–470.
6. Sheridan RP, Feuston BP, Maiorov VN, Kearsley S: **Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR.** *J Chem Inf Comput Sci* 2004, **44**:1912–1928.
7. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R: **Comparison of different approaches to define the applicability domain of QSAR models.** *Molecules* 2012, **17**:4791–4810.
8. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MT, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D, Schultz T, Stanton DW, van de Sandt JJ, Tong W, Veith G, Yang C: **Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52.** *Altern Lab Anim* 2005, **33**:155–173.
9. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T: **QSAR applicabilty domain estimation by projection of the training set descriptor space: A review.** *Altern Lab Anim* 2005, **33**:445–459.
10. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan OA: **Stepwise approach for defining the applicability domain of SAR and QSAR models.** *J Chem Inf Model* 2005, **45**:839–849.
11. Breiman L, Meisel W, Purcell E: **Variable kernel estimates of multivariate densities.** *Technometrics* 1977, **19**:135–144.
12. Asikainen AH, Ruuskanen J, Tuppurainen KA: **Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands.** *Environ Sci Technol* 2004, **38**:6724–6729.
13. Tropsha A: **Best practices for QSAR model development, validation, and exploitation.** *Mol Inf* 2010, **29**:476–488.
14. Cedeño W, Agrafiotis DK: **Using particle swarms for the development of QSAR models based on K-nearest neighbor and kernel regression.** *J Comput Aided Mol Des* 2003, **17**:255–263.
15. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, 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.
16. Itskowitz P, Tropsha A: **k nearest neighbors QSAR modeling as a variational problem: theory and applications.** *J Chem Inf Model* 2005, **45**:777–785.
17. Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell JB: **Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization.** *J Chem Inf Model* 2006, **46**:2412–2422.
18. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A: **Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection.** *J Chem Inf Model* 2008, **48**:1733–1746.
19. Silverman BW: *Density Estimation for Statistics and Data Analysis.* London, UK: Chapman and Hall; 1986.
20. Zhao C, Boriani E, Chana A, Roncaglioni A, Benfenati E: **A new hybrid system of QSAR models for predicting bioconcentration factors (BCF).** *Chemosphere* 2008, **73**:1701–1707.
21. Lombardo A, Roncaglioni A, Boriani E, Milan C, Benfenati E: **Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish.** *Chem Cent J* 2010, **4**(Suppl 1):S1. doi:10.1186/1752-153X-4-S1-S1.
22. Chambers JM, Cleveland WS, Kleiner B, Tukey PA: *Graphical methods for Data Analysis.* Pacific Grove, CA: Wadsworth & Brooks/Cole; 1983.
23. *Box plot – MATLAB.* http://www.mathworks.it/it/help/stats/boxplot.html.
24. Consonni V, Ballabio D, Todeschini R: **Comments on the definition of the Q2 parameter for QSAR validation.** *J Chem Inf Model* 2009, **49**:1669–1678.
25. Consonni V, Ballabio D, Todeschini R: **Evaluation of model predictive ability by external validation techniques.** *J Chemometr* 2010, **24**:194–201.
26. *DRAGON (Software for Molecular Descriptor Calculations). Talete srl, Milano, Italy.* http://www.talete.mi.it.
27. Forina M, Lanteri S, Armanino C, Cerrato Oliveros C, Casolino C: *V-PARVUS software, User manual.* 2004 http://parvus@difar.unige.it.
28. *MATLAB. The Language of Technical Computing.* http://www.mathworks.com/products/matlab/.
29. Todeschini R, Ballabio D, Consonni V, Sahigara F, Filzmoser P: **Locally-centred Mahalanobis distance: a new distance measure with salient features towards outlier detection.** *Anal Chim Acta* 2013. doi:10.1016/j.aca.2013.04.034.