Molecular Descriptors
the free online resource

Defining the Applicability Domain of QSAR models: An overview

Faizan Sahigara

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1-20126 Milano, Italy.

What is Applicability Domain ?

QSARs establish a quantitative relationship between chemical structures and their properties [1]. In theory, QSAR models can be used to predict the properties of chemical structures, provided their structural information is available. In the recent years, there had been a growing awareness about QSARs and their applications. This is quite evident from their use for regulatory purposes. A new European legislation on chemicals – REACH (Registration, Evaluation, Authorization and restriction of Chemicals) came into force in 2007, allows and encourages the use of QSAR model predictions when the experimental data are not sufficiently available or as supplementary information, provided validity of the model is justified [2,3].

However, this rising popularity of QSAR models is also accompanied by a question over their reliable predictions [4]. In theory, derivation of QSAR models is based primarily on training sets which are structurally limited and thus, their applicability to the query chemicals is limited [5]. Thus, their applicability towards reliable predictions is restricted in a chemical space to some specific chemical categories. Such reliable predictions are usually confined to those chemicals, that are structurally similar to the training compounds used to build the model [6-8].

The principle of Applicability Domain obliges the users to specify the scope of their proposed models thus, defining the model limitations with respect to its structural domain and response space. If an external compound is beyond the defined scope of a given model, it is considered outside that model's Applicability Domain (AD) and cannot be associated with a reliable prediction.

What are the key aspects in defining the AD of QSAR models ?

1) Identification of the subspace of chemical structures that can be predicted reliably.
2) Defined AD determines the degree of generalization of a given predictive model. Thus, if the AD is too restricted, it implies the model can provide reliable predictions for very limited chemical categories.

3) A well defined AD indicates if the endpoint for the chemical structures under evaluation can be reliably predicted.
4) Characterization of the interpolation space is very significant to define the AD for a given QSAR model.

How can the AD of a model be defined ?

Several strategies towards defining the Applicability Domain of QSAR models have been proposed in literature. This section of the tutorial aims to provide an overview of some major AD approaches.

1) Range-based Methods

a) Bounding Box

Considering the range of individual descriptors used to build the model, this approach defines the AD as a Bounding Box which is a *p*-dimensional hyper-rectangle defined on the basis of maximum and minimum values of each descriptor used to build the model.

Drawbacks: Empty regions in the interpolation space cannot be identified and also the correlation between descriptors cannot be taken into account [1,4].

b) PCA Bounding Box

Similar to the earlier approach, however, the AD is defined considering the projection of the molecules in the principal component space and taking into account the maximum and minimum values for the PC scores. This approach resolves the problem of correlation between descriptors.

Drawbacks: Empty regions within the interpolation space still cannot be identified [1,3-4].

2) Geometric Method

This approach characterizes the interpolation space by defining a smallest convex area containing the entire training set.

Drawbacks: Increasing data complexity highly affects the implementation of a convex hull. For a data with two or three dimensions, the method works efficiently however, with further increase in dimensions, the implementation adds to the complexity of the algorithm [1,9]. Set boundaries are analyzed without considering the actual data distribution. Convex Hull cannot identify the potential internal empty regions within the interpolation space [1,2].

3) Distance-based Methods

These approaches define the AD by calculating distances of a query compound from a defined point within the descriptor space of the training data. This measured

distance between defined point and the dataset is then compared with a pre-defined threshold. However, no strict rules are evident from the literature about this pre-defined threshold and thus, it is up to the user to take an appropriate decision towards defining the thresholds [1-5].

Some commonly used and most useful distance measures in QSAR studies include Mahalanobis, Euclidean and City Block distances. Leverage is another measure that is recommended in defining model's AD [10]. In theory, Leverage is proportional to Hotellings T^2 statistic and Mahalanobis distance measure from the centroid of the training set [4]. Usually, a warning threshold is set to three times the average of the leverage *p/n*, where *p* is the number of model parameters while *n* is the number of training compounds. Query compounds with leverage higher than this defined threshold of 3**p/n* are considered to be unreliably predicted.

Drawbacks: Lack of strict rules in literature towards defining the thresholds can lead to ambiguous results. Correlated descriptors can be handled using Mahalanobis distance or Leverage, since they use co-variance matrix for their calculations, however, an additional treatment like PC rotation is required for other distance measures.

4) Probability Density Distribution based Methods

These approaches defines a model's AD by estimating the Probability Density Function for the given data. The estimation of Probability Density Function is feasible by both, parametric methods that assume standard distribution and non parametric methods which do not have any such assumptions concerning the data distribution. These approaches are considered to be efficient due to their ability to identify the internal empty regions and reflecting actual data distribution by generating concave regions around the extremities of the interpolation space [1,4].

Potential function is calculated for all the training compounds, followed by which a global potential is obtained by adding the individual potentials, thus indicating the potential density [11,12]. A percentile value for the probability density is opted and a threshold value is defined. Finally, those query compounds having potential function values lower than this threshold are considered to be outside the AD.

5) K Nearest Neighbours Approach

This approach defines the model's AD by assessing the similarity between training and test compounds. Distance of a query compound from its nearest training neighbour or its average distance from k nearest training neighbours is calculated. The calculated distances for test compounds are then compared with a pre-defined threshold. The test compounds with low distance to the training set is associated with higher number of training compounds and thus, is considered to be reliably predicted [8].

6) Decision Trees and Decision Forests Approach

This approach defines the AD for a consensus prediction of Decision Trees, in terms of prediction confidence and domain extrapolation. Decision Trees are combined and

the distance between them are kept to maximum possible, thus to minimize the overfitting. Prediction confidence for a given compound is determined by averaging the predictions derived from all combined Decision Trees while, its prediction accuracy outside the training space is represented by the domain extrapolation [1,13,14].

7) Stepwise Approach to Determine Model's AD

With the Stepwise approach, AD of a QSAR model is better assessed executing four stages in a sequential manner. First stage checks if a test compound is within the variation range of physicochemical properties of the training compounds. Next, a structural similarity check is made for those compounds that were reliably predicted by the model. A mechanistic check is made in the third stage while, the last stage takes into account the reliability of simulated metabolism [3,5].

8) Distance to Model Approach

Distance to Model (DM) approach [15] estimates the prediction quality by using the information about the target property. Thus, the information about prediction itself is used for AD evaluation.

Standard Deviation (STD) DM: This method uses the standard deviation of predictions vector as the DM, since for given predictions from different set of models based on the same data, significant discrepancy in values indicates unreliability of a prediction.

CORREL: This method is derived from ensemble of models and is based on correlation of vectors of ensemble's predictions for the target and training set compounds. Compounds are considered to be 'near to the model' if they have a higher value for correlation coefficient.

CLASS-LAG: Prediction accuracy for classification models is provided by the CLASS-LAG measure that signals the confidence in prediction based on the idea that values closest to the classification label {+1,-1} are more reliably predicted and those that are nearer to 0 can be associated with 'uncertainty area' indicating an unreliable prediction [16].

PROB-STD: This DM combines the information from STD and CLASS-LAG. Lowering of STD value and approach towards the classification label {+1,-1} results in lower value for PROB-STD, indicating a reliable prediction.

9) Virtual models for property Evaluation of chemicals within a Global Architecture (VEGA)

VEGA [17] uses Applicability Domain Index (ADI) as a major criterion in defining Applicability Domain of predictions. The values for this Index ranges from 0 (worst case) to 1 (best case). In theory, this index is derived evaluating several other indices, each of which focuses on a specific aspect relevant in defining the AD. The values for each index including the main ADI is categorised into three different intervals to indicate if the evaluation was positive, suspicious or negative.

Following are the components reported in VEGA platform for AD assessment:

1) Similar molecules with known experimental value
2) Accuracy (average error) of prediction for similar molecules
3) Concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules)
4) Maximum error of prediction among similar molecules
5) Atom Centered Fragments similarity check
6) Descriptors noise sensitivity analysis
7) Model descriptors range check and
8) Global AD Index.

Further Reading

Recently, following publication was made available online discussing the results derived from several classical descriptor-based AD approaches on existing validated datasets from the literature:

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.

References

1. Netzeva, T.I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; *et al.* 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.

2.REACH - European Community Regulation on chemicals and their safe use. Available online: http://ec.europa.eu/environment/chemicals/reach/ reach_intro.htm

3. Worth, A.P.; Bassan, A.; Gallegos, A.; Netzeva, T.I.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vracko, M. *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*. ECB Report EUR 21866 EN, European Commission, Joint Research Centre; Ispra, Italy, 2005; p. 95.

4. 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.

5. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O.A. Stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–49.

6. Worth, A.P.; van Leeuwen, C.J.; 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.

7. 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.

8. Sheridan, R.; Feuston, R.P.; Maiorov, V.N.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1912–1928.

9. Preparata, F.P.; Shamos, M.I. Convex hulls: Basic Algorithms. In *Computational Geometry: An Introduction*; Preparata, F.P., Shamos, M.I., Eds.; Springer-Verlag: New York, NY, USA, 1991; pp. 95–148.

10. Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models. *QSAR & Comb. Sci.* **2003**, *22*, 69–77.

11. Jouan-Rimbaud, D.; Bouveresse, E.; Massart, D.L.; de Noord O.E. Detection of prediction outliers and inliers in multivariate calibration. *Anal. Chim. Acta* **1999**, *388*, 283–301.

12. Forina, M.; Armanino, C.; Leardi R.; Drava, G. A class-modelling technique based on potential functions. *J. Chemometr.* **1991**, *5*, 435–453.

13. Tong, W.; Hong, H.; Fang, H.; Xie, Q. Perkins, R. Decision forest: Combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525–531.

14. Tong, W.; Hong, H.; Xie, Q.; Xie, L.; Fang, H.; Perkins, R. Assessing QSAR limitations: A regulatory perspective. *Curr. Comput. Aid. Drug*. **2004**, *1*, 65–72.

15. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Kovalishyn, V.V.; Prokopenko, V.V.; Tetko, I.V. Applicability domain for in silico models to achieve accuracy of experimental measurements J. Chemometrics., 2010, 24(3-4), 202-208.

16. Manallack DT, Tehan BG, Gancia E, Hudson BD, Ford MG, Livingstone DJ, Whitley DC, Pitt WR. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. J. Chem. Inf. Comput. Sci. 2003; 43(2): 674–679.

17. http://www.vega-qsar.eu/index.php