

LIFE Project Acronym and Number

ANTARES
LIFE08 ENV/IT/000435

Deliverable Report

Deliverable Name and Number

Deliverable 2
**Report on the identified criteria for non-testing methods,
and their scores**

Deliverable Date

30/09/2010

Deliverable Data

Associated action	Action 2: Identification of the criteria for the non-testing methods for the REACH legislation
Beneficiary Responsible	Istituto di Ricerche Farmacologiche "Mario Negri"
Contact person	Dr. Emilio Benfenati
Contributors	Dr. Chiara Milan, Dr Alessandra Roncaglioni, Dr Rodolfo Gonella Diaza, Dr Antonio Cassano, Dr. Nazanin Golbamaki, Dr Emilio Benfenati (external: Dr Alberto Manganaro, for applicability domain criteria)

SUMMARY

The REACH legislation promotes the use of non-testing methods (NTM), which are all the approaches used to predict the effects of chemical compounds without the use of the real chemical compound, but on the basis of the chemical structure only and comprise a series of different tools whose commonality lies in the identification of a relation between chemical structure and exhibited activity or toxicity.

There are different types of NTM to assess activity and/or toxicity of chemicals, among them Quantitative structure-activity relationship (QSAR) methods are the most used. QSAR is the process by which chemical structure is quantitatively correlated with a well-defined process, such as biological activity or chemical reactivity.

As the project has the main target to analyze the use of NTM in accordance to REACH, and to identify suitable method, the first step has been to identify relevant criteria for comparing QSAR methods.

A list of them has been established. They have been divided into main (more important) and additional criteria. For each of them a particular score has been decided based on the importance of the criterion.

Using the criteria here identified, in the following Actions we will first identify available QSAR models which could be used for REACH. Then, if there is more than one method for the same endpoint, we will rank them, using the proposed scoring system.

This deliverable contains the main results associated to the Action 2 of ANTARES project.

1. INTRODUCTION

The term Non-Testing Methods (NTM) refers to all the approaches used to predict the effects of chemical compounds without the use of the real chemical compound, but only on the basis of the chemical structure and comprises a series of different tools whose commonality lies in the identification of a relation between chemical structure and exhibited activity or toxicity.

Different types of NTM to assess activity and/or toxicity of chemicals can be enumerated:

- QSAR: stands for Quantitative Structure-Activity Relationship; it is a mathematical model relating one or more quantitative parameters, which are derived from the chemical structure, to a quantitative (in case of regression models) or qualitative (in case of classification models) measure of a property or activity.
- SAR: stands for Structure-Activity Relationship; is a qualitative relationship that relates a (sub)structure to the presence or absence of a property or activity of interest
- Read-Across: is a method to fill data-gap where endpoint information from one or more chemicals is used to predict the same endpoint for another chemical which is considered to be similar. The similarity allows identifying a group or a category of compounds where physicochemical and/or toxicological and/or ecotoxicological properties are likely to be similar or follow a regular pattern.
- Grouping: in this case a group of compounds which share a similar structure, are addressed as a family, supporting the assignments of missing property values on the basis of the knowledge of the values of the other members of the family.

Although from a purely scientific perspective this field of research originates decades ago, discussion is ongoing about requirements of these methods for being accepted for regulatory purposes.

In this context ANTARES aims to evaluate existing NTM and their feasibility for REACH purposes and as a first step in this process the relevant criteria for judging and comparing these different methods.

2. EXAMINATION OF SOURCES FOR CRITERIA IDENTIFICATION

REACH criteria for QSARs

As the project has the main target to analyze the use of NTM in accordance to REACH the first source of information to identify relevant criteria to consider QSAR estimations useful for REACH was found directly in the sentences reported in the EU law [in particular annex XI, 1.3, 1.4, 1.5] in which are considered Qualitative or Quantitative structure-activity relationship ((Q)SAR), grouping of substances and read-across approach. Below, we will refer to QSAR, including both the methods, QSAR and SAR.

We specify that QSAR models for regulatory purposes can be used for at least three purposes: registration, classification, and prioritisation. The requirements for these purposes are very different, and in particular the acceptable accuracy of the results.

Results obtained from suitable QSAR models may indicate the presence of a certain dangerous property or may be important in relation to the property understanding, which may be important for the assessment.

About grouping of substances and read-across approach, substances whose physico-chemical and ecotoxicological properties are likely to be similar or follow a regular pattern as a result of structural similarity may be considered as a group, or 'category' of substances. Application of the group concept requires that physico-chemical properties, human effects and environmental effects or environmental fate may be predicted from data for reference substances within the group by interpolation to the other substances in the group (read-across approach). This avoids the need to test every substance for every endpoint.

More in particular for (Q)SAR this is what REACH says (Annex XI):

"Results of (Q)SARs may be used instead of testing when the following conditions are met:

- results are derived from a (Q)SAR model whose scientific validity has been established,*
- the substance falls within the applicability domain of the (Q)SAR model,*
- results are adequate for the purpose of classification and labelling and/or risk assessment,*
- adequate and reliable documentation of the applied method is provided."*

Let's analyse these requirements:

- results are derived from a (Q)SAR model whose scientific validity has been established,*

The OECD adopted in 2004 five principles for establishing the validity of (Q)SAR models for use in regulatory assessment of chemical safety (<http://www.oecd.org/dataoecd/33/37/37849783.pdf> OECD principles for the validation, for the regulatory purposes, of (Quantitative) Structure-Activity Relationship models (2004).

In particular for the evaluation the following aspects should be addressed:

- 1) a defined endpoint
- 2) an unambiguous algorithm
- 3) a defined domain of applicability
- 4) appropriate measures of goodness-of-fit, robustness and predictivity
- 5) a mechanistic interpretation, if possible

It has to be pointed out that still these OECD principles are quite general; as an example no suggestions are given about specific parameters requested to evaluate the robustness of a method, nor precise indications

about appropriate methods to define the applicability domain. The formerly ECB institute of JRC developed a more detailed checklist which has been formalized in the (Q)SAR Model Reporting Format (QMRF).

Some specific criteria may apply for the specific situations. In particular for REACH the endpoint has to be defined and should belong to the REACH endpoint list. It is important to define the experimental protocol and consider the endpoint data quality and its variability because of the experimental data derive from different sources. The intent is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions.

It is important to explicit the definition of the algorithm, the descriptors in the model and its variability.

It is also necessary to know the software name and version for the descriptor generation and the chemical format to use.

It is necessary to specify the method used to assess the applicability domain and the software name and the version used and to underline the limits of the applicability domain.

For the internal validation it is necessary to know if the training test and what kind of information about it, are available, and to explicit which kind of data are used for each descriptor variable and for the dependent variable for the training set.

We need appropriate measures of goodness-of-fit, robustness and predictivity. It is important to verify both the internal performance of a model (as represented by goodness-of-fit and robustness) and the predictivity of a model (as determined by external validation).

It may be useful to provide a mechanistic interpretation if possible and consider *a priori* or *a posteriori* mechanistic interpretation.

The preliminary document on (Q)SAR characterisation, compiled by the formerly known European Chemicals Bureau (ECB), lists a series of statistical parameters to be used for the model evaluation. Different tools apply to a model which is a classifier, or to a model which is a regression method. In the first case the output of the model is a class or category, such as toxic, or mutagen.

Evaluation of a classifier

Most typically classifiers are evaluated using the Cooper statistic. In the simple case of a binary classification, there are two classes, such as toxic (positive) or not (negative). The results of a classifier could be therefore grouped in four cases: toxic compounds predicted as toxic (True Positive or TP) or as non toxic (False Negative or FN) as well as non toxic compounds predicted as non toxic (true negative or

TN) or as toxic (False Positive or FP). Three main statistical parameters can be derived by the combination of these four cases, for the model evaluation:

- **Accuracy (A)**, also referred as concordance, is the measure of the correctness of prediction. This parameter gives a general evaluation of the errors done and is defined as the ratio between the compounds correctly predicted to the total number of compounds. Good models have high accuracy value.

$$A = (TP + TN) / \text{TOTAL}$$

- **Sensitivity (S)** is the measure of the positive compounds correctly predicted. Especially for regulatory purposes, it is important not to declare safe a chemical which conversely it is toxic (FN). The sensitivity takes into account the number of FN and is defined as the ratio of the TP tests to the total number of positives. A good model has high sensitivity.

$$S = TP / P$$

- **Specificity (SP)** is the measure of the negative compounds correctly predicted. Specificity keeps into account the number of false positives and is defined as the ratio of the TN tests to the total number of negative compounds. Sometimes the **1 - SP** parameter is reported.

$$SP = TN / N$$

It is our opinion that for regulatory purposes it is important to verify that the classifier has a high sensitivity, in order to reduce the number of false negatives.

Not only binary classifications are defined within REACH. For instance, a chemical can be not bioaccumulative, or bioaccumulative, or very bioaccumulative (three classes).

Evaluation of a regression model

Regression models are most typically evaluated using statistical parameters which keep into account the errors of the model. These errors are measured on the basis of the training set, and this gives an idea of the model robustness. However, this is not sufficient since the main interest of REACH is to understand if a certain model can be used for prediction purposes. Thus, for regulatory purposes additional statistical measurements are used, for predictivity. Some measurements use internal validation, other tools refer to an external test set.

The values predicted by the model (on training, test and/or external validation set) are put in correlation with the experimental values using a graph (an example is shown below) and the **coefficient of**

determination (R^2) is calculated and gives an estimation of the model goodness. (<http://www.orchestra-qsar.eu>)

- the substance falls within the applicability domain of the (Q)SAR model,

The need to define an applicability domain expresses the fact that (Q)SARs are models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions.

It is important to define the applicability domain, in particular to describe the applicability domain of the model and, for this reason, it is useful to characterize the reasons why a compound is in or out a particular AD and to specify the chemical class for the compounds out of domain. Find attached the issue of the Applicability Domain.

- results are adequate for the purpose of classification and labelling and/or risk assessment

Results must be clear: no confusion about the results' nature and the characteristics. The project purpose is clearly defined and described, the achievement have been evaluated in regards to the purpose. The results' nature, features, *pros* and *cons* are clearly detailed. They should refer to the legislation, in order to be adequate. In this point, the legislation states that its interest and evaluation refers to the purpose of REACH, and the model can be used for these two different reasons: risk assessment and/or classifications and labelling. We notice that in the first case the value should be a continuous one, while in the second case a categorical output is appropriate.

- adequate and reliable documentation of the applied method is provided."

This criterion requires transparency since all documentation at the basis of the assessment of the properties of a chemical should be clearly available and checkable. One of the driving forces of REACH was to have the correct knowledge on the properties of the chemical substances on the market.

Besides the criteria specified by REACH there are other additional criteria:

The **reproducibility**: in particular, what kind of factors are involved in reproducibility. The reproducibility is quite important, if we want to compare results obtained in different countries or provided by different stakeholders. If we imagine a method which is likely to give different results, on the basis of the user, this may easily become criticised for the specific application: application for regulatory purposes.

The **easiness** of the model is related to this issue. If we imagine a model which is complicated, and has several parameters to be chosen, we may easily get different results. The outcomes of the project should respond to users' needs.

The **clarity** of the result should be another criterion, as the comprehension of the results. It may happen that the output of the model is of difficult interpretation. The project purpose has to be clearly defined and described and the achievements have to be evaluated in regards to the purpose. The results' nature, features, *pros* and *cons* are clearly detailed.

The **access** to the model is another criterion: some models are free, others very expensive.

The **time** necessary to get the results (speed of the model) is another desirable criterion.

Another useful feature is the possibility to run prediction **in batch** , in order to save time.

3. The issue of Applicability Domain.

3.1. Introduction

One of the relevant issue to consider in order to evaluate the reliability of QSAR models is the assessment of the prediction reliability, i.e. the study of the Applicability Domain (AD) of the model.

The AD consists of the chemical space within which predictions for a given model can be considered reliable. This means that if a compound falls within the AD, its prediction should be reliable, but if it lays outside the AD its prediction can be an extrapolation, thus being unreliable. The importance of having a well-defined AD in order to validate a QSAR model has been pointed out by the OECD in its principles (OECD principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship models, <http://www.oecd.org/dataoecd/33/37/37849783.pdf>).

As it can be seen the assessment of the AD strictly depends upon the original training set used to build the model. In general, all approaches to assess the AD are related to a measurement of how a compound to be predicted is far from the training set molecule. Thus, from the compounds in the training set can be deduced what is the chemical space where the model should result reliable. From this overview comes that the issues to be faced are:

1. The choice of the features to be taken into account to define the model's chemical space
2. The mathematical methods to be used to calculate whether a compound falls into the AD or not.

Furthermore, the study of the AD can be extended if there is the possibility to calculate and to associate a measurement of the unreliability when a compound falls out of the AD.

3.2. Model's space definition

There are different approaches for the choice of what features would be used to define the model space. The chemical space should ideally include the structural, physico-chemical and response space of the

model. This is because the best assurance that a chemical is predicted reliably is to have confirmation that the chemical is not an outlier in terms of its structural fragments (structural domain), its descriptor values (physicochemical domain) or its response values (response domain).

3.2.1. Physico-chemical domain

One of the most common choices for designing the chemical space is to use the molecular descriptors of the model. Those descriptors are the ones that have been selected to build the QSAR model, thus they have a strong connection with the studied chemical property/activity. Usually all descriptors are taken into account, in some cases greater relevance can be given to some particular descriptors if their particular role in the model is well known.

3.2.2 Structural domain

Besides the original model descriptors, a kind of chemical structural information can be considered. In this case there are two different paths that can be followed.

A generic structural approach will consider some encoding of the chemical structure of all compounds into the training set. For example, fingerprints are often used for this purpose, as they are quite easy to calculate and manipulate. In this case no particular information coming from the model is considered, as the space is designed by the structure of all training set compounds.

Often models are built with particular attention to the mode of action (MOA) thus a model contains some mechanistic information i.e. explicit link between some fragments or topologic feature and the studied activity/property. In this case the model space can be also defined by the presence or absence of certain fragments.

3.2.3 Response domain

Finally, the model response can be taken into account.

3.3. Model's space calculation

Once the features to be used for the AD are defined, there is a wide choice of approaches to process the model's space and to calculate whether a new compound should be considered inside the space or not.

3.3.1 Calculation in Physico-chemical domain

Probably the simplest method is to work on ranges. For instance, if only the molecular descriptors are considered, the model's space could be defined with the range of all descriptors; a new compound will be into the AD if all its descriptors have values between the minimum and maximum values of the model's descriptors (Jaworska, J., Nikolova-Jeliazkova, N. and Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: a review, Altern Lab Anim 33, 445-459).

Most of the approaches start from a similar idea, but try to refine the definition of the space as of course using ranges leads to a very rough definition. An improvement of the ranges approach is the application of

a Principal Component Analysis (Nikolova-Jeliazkova, N. and Jaworska, J. (2005). An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN, Altern Lab Anim 33, 461-470).

Another approach is based on the calculation of a measurement of distance. In this case, first a distance formula should be chosen (frequently used distances are Euclidean, Mahalanobis, City Block). After the calculation of the distance of a new compound towards all training set compounds, an algorithm to define whether it is or not into the AD should be defined. For instance, it could be considered the maximum distance, or the average value, and then set thresholds to define the model's space.

Other kind of distance that could be used are Hotelling T2 test and leverage (Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., McDowell, R. M. and Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ Health Perspect 111, 1361—1375).

A further improvement for refine the definition of the space is the use of probability density distribution (both parametric and non-parametric) (Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. 176pp. Boca Raton, FL, USA: CRC Press).

This technique can detect better empty model's space regions thus bringing a better AD evaluation.

3.3.2 Calculation on structural domain

In the structural domain, two different approaches can be considered.

A more generic approach chooses a method of measurement of structural similarity, then applies it to calculate the distance of the new compound towards the training set compounds. Most common techniques to calculate structural similarity are based on fingerprinting (Willett, P. (2011). Similarity searching using 2D structural fingerprints, Methods Mol Biol 672, 133-158).

Fingerprints represent a good trade-off between the complexity of molecule structure encoding and the simplicity in storing them and making calculation.

A slightly different method for considering structural similarity, consisting in the detection and comparison of presence/absence of atom-centered fragments, has been tested (Kühne, R., Ebert, R.-U. and Schüürmann, G. (2009). Chemical domain of QSAR models from atom-centered fragments, J Chem Inf Model 49, 2660-2669).

Another technique uses SMILES attributes, i.e. fragments extracted from the SMILES representation of the molecule, still based on the comparison of presence/absence of these features to evaluate the belonging to

the AD (Toropov, A. A., Toropova, A. P., Benfenati, E. and Manganaro, A. (2009). QSAR modelling of carcinogenicity by balance of correlations, *Mol Divers* 13, 367-373).

Those methods are generic, thus they can be applied to every model where the training set is available. A different approach consists in the study of the single QSAR model in order to obtain some structural information coming from chemical theory and/or statistical analysis of the model. In this case, some structural features (i.e. fragments) are identified and used to define the space of the model or to define outliers.

These techniques are often useful with QSAR models based on structural features and knowledge-based rules (like well-spread programs, DEREK and MULTICASE). But in general, they can be useful when modeling activities for which some mechanistic action is known; this is the case, for example, of skin sensitization models (Aptula, A. O., Patlewicz, G. and Roberts, D. W. (2005). Skin sensitization: reaction mechanistic applicability domains for structure-activity relationships, *Chem Res Toxicol* 18, 1420-1426). This approach can be used also in other kind of models. For example, in the CAESAR BCF model (Lombardo, A., Roncaglioni, A., Boriani, E., Milan, C. and Benfenati, E. (2010). Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish, *Chem Cent J* 4 Suppl 1, S1).

Outliers have been studied, and some fragments that identify their chemical classes have been found. Thus, the presence of one of these fragments in a new compound leads to its placement outside the AD.

3.3.3 Calculation on response domain

Upon response domain, approaches similar to those used in physico-chemical domain can be used. In general, the underlying concept is that values outside the space defined by training set's responses could point out an outlier compound.

For example, in the CAESAR version 2.0 models (CAESAR project website: <http://www.caesar-project.eu/>) a composite AD check is implemented, that takes into account information from all mentioned domains. Concerning the response, experimental values from the most similar compounds found in the training set are considered and confronted with the new compound's prediction. Thus, if a compound prediction falls in a response space region where experimental values are heterogeneous, or if the prediction disagrees with the common value found in the response space region, the compound is marked as possibly out of AD.

4. IDENTIFICATION OF MAIN CRITERIA

Specific considerations for read-across

In case of Read-across it is difficult to judge the method itself because two key aspects that may influence the reliability of the obtained prediction are the quality of existing data used in the extrapolation and the consistency in the group definition/identification.

For Read-Across and grouping the correctness of the values is much more critical than in the case of QSAR. Furthermore, a major role is left to human decision and no formalisation has been defined, as in the case of QSAR. Those two factors introduce a large amount of uncertainty, and reproducibility is critical. The main factors we addressed for QSAR will remain but in a simplified way. Obviously the key element is that the assessment is correct. Keeping into account the uncertainty of variability of the results as discussed, it is likely that due attention has to be given to this through some tests for repeatability and reproducibility of the results within the same and different laboratories.

5. DEFINITION OF THE SCORING SYSTEM APPLIED TO THE SELECTED CRITERIA

On the basis of all the above-described points, we identified a set of criteria. Some are more important, other less. Thus, we assigned different scores to these criteria. In the table below we list the main criteria, which should ideally be covered for all models. If a model gets a score of 0 for one of the following criteria, this is a critical issue.

Table 1. List of the main criteria used within ANTARES to select the QSAR models.

<u>NAME</u>	<u>DESCRIPTION</u>	<u>SCORE</u>
1. Data Quality	<p>We will consider if the data set is the best quality for a specific endpoint; then an additional criterion will be number of compounds. Of course only endpoints useful for REACH will be addressed within ANTARES.</p> <p>This is based on the relevance and the quality score.</p> <p>Relevance: 0, 1 borderline; 2 exact.</p> <p>Quality: 0 no info; 1; good quality. The second subcriterion is applied only if the data are relevant for REACH, thus only if the first subcriterion is > 0.</p>	0-3
2. Chemical number	<p>We will consider the number of compounds.</p> <p>0 = < 100</p> <p>1 = 100 - 500</p> <p>2 = 500 - 5000</p> <p>3 = > 5000</p>	0-3
3.Descriptors/fragments	<p>We will give preference to models where explicit descriptors/fragments are defined; including software name and version for the descriptor generation.</p> <p>3 = full description, equation available</p> <p>2 = possible ambiguities depending on the chemical format</p> <p>1 = only partial info available</p> <p>0 = no info</p>	0-3
4. Explicit and verified the algorithm	<p>We will give preference to models where explicit algorithm is defined; Defining the algorithm: explicit algorithm.</p>	0-3

	<p>3 = full description, equation available</p> <p>2 = possible ambiguities</p> <p>1 = only partial info available</p> <p>0 = no info</p>	
5. Applicability domain	<p>Description of the AD of the model. Method used to assess the AD. Software name and version for AD assessment. Limits of AD.</p> <p>3 = full description, and model provides tool</p> <p>2 = explained, but to be applied manually</p> <p>1 = only partial info available</p> <p>0 = no info</p>	0-3
6. Performance	<p>We refer to the Statistical description provided. We adopted the following thresholds for R²: 2 > 0.85</p> <p>1 = 0.85 – 0.65</p> <p>0 < 0.65</p> <p>In addition if the training set is available: 1</p>	0-3
7. Validation	<p>We refer to the internal validation: availability of the training set; available information for the training set; data for each descriptor variable for the training set; data for the dependent variable for the training set; statistics for goodness-of-fit.</p> <p>We adopted the following thresholds for Q²</p> <p>2 > 0.80</p> <p>1 = 0.60 – 0.80</p> <p>0 < 0.60</p> <p>In addition if the external validation is available: 1</p>	0-3
8. Output	<p>To explicit its format, to verify it and to test the comprehension for the user. We will give preference to models where input is clearly defined, and how to use the results is explicated. As further criteria we will evaluate if the usability has been checked.</p> <p>It is based on three components, each scored 0 or 1.</p> <p>Univocous</p> <p>Usable as it is</p> <p>Usable as key study or not</p>	0-3

9. Cost	<p>We will give preference to models which are less expensive.</p> <p>3 = free</p> <p>2 = perpetual license</p> <p>0 = annual license and cost</p>	0-3
---------	--	-----

The best model is that that addresses the seven aspects with the maximum score.

Other criteria will be also used for a more detailed evaluation, in case there are more than one model available for the same endpoint. In case of equal scores the comparison will be done also with the additional criteria, shown in the table below.

Table 2. List of the additional criteria used within ANTARES to select the QSAR models.

<u>NAME</u>	<u>DESCRIPTION</u>	<u>SCORE</u>
10. Batch supported	If it is possible to calculate properties of a set of chemicals	0-2
11. Structure format	We will give preference to models where an explicit structure format is defined and thus as further criteria if it has been check independently.	0-1
12. Verify the presence of the uncertainty	We will give preference to models which address uncertainty, and as further criterion if this is specific for a certain compound.	0-1
13. Further adequate and reliable documentation	We will give preference to models where this is available	0-1
14. Usability/user friendly	We will give preference to models which are easier to be used	0-1
15. Comprehension	We will give preference to models which are easier to comprehend	0-1
16. Skill requested to interpret results	We will give preference to models which are easier to be interpreted	0-1
17. Access	We will give preference to models which have a better access	0-1
18. Platform/software requirements	We will give preference to models which have less requirements	0-1
19. Connection problems	We will give preference to models which have a less problems in	0-1

	connections	
20. Time needed	We will give preference to models which are faster	0-1

Thus, these are the overall codified criteria which will be used, first to identify models which have the main criteria satisfied, and in case of more models available, we will use the whole set of criteria to identify the most useful ones for our purposes.