# Read across strategy for the assessment of dyes with the aid of a new QSAR system

V. Alberti[a], C. Rovida[a], D. Ballabio[b], V. Consonni[b], M. Locatelli[a], R. Todeschini[b] and M. Kahlberg[a]

a) REACH&Colours Italia srl, via Locatelli 6, 20124 Milano, Italy

b) Milano Chemometrics and QSAR Research Group, Environmental Science Department, University of Milano-Bicocca, 20126 Milano, Italy

Corresponding address: vanessa.alberti@reachcolours.it

## 1. DYES and REACH Regulation

Colorants are listed according to the widely acclaimed system of Colour Index Generic Names and Colour Index Constitution Numbers. The whole Color Index database contains about 1500 dyes which are complex molecules: high molecular weight, different organic functionalities (azo dyes, metal complex dyes etc), multi component substances and different counter ions.

The risk assessment is particularly sensitive due to the wide spread use and consumer exposures. The number of requested data to comply with REACH Regulation would be huge and unaffordable from an economical, timing and animal welfare point of view.

## 2. REACH&Colours Italia Srl

REACH&Colours Italia Srl is the manager of the three European consortia (paper, textile and leather) born for the registration of dyes.

Further information on ETAD (The Ecological and Toxicological Association of Dyes and Organic Pigments Manufacturers) website (http://www.etad.com/index.php Highlights section, REACH Dyes Consortia).

Companies participating in consortia gave the possibility to use all proprietary data that have been performed on dyes in the last 30 years.

## 3. UNIMIB

The Milano Chemometrics and QSAR Research Group (MCQ), coordinated by Prof. Todeschini, has been established in the UNIMIB and has more than 25 years experience in chemometrics, QSAR, molecular descriptors, ANNs, multicriteria decision making, computer programming, and experimental design.

In the latest years, MCQ was very active in the research field of molecular modeling proposing new molecular descriptors. Some important products of their research are: the software DRAGON for the calculation of molecular descriptors; several MATLAB toolboxes for multivariate data analysis; the proposal of novel strategy to define QSAR model applicability domain and new models for predicting BCF and biodegradability of substances.

Since 2009, their activities have also been focused on QSAR and chemometrics topics of relevance to REACH.

## 4. QSAR for dyes

As part of the present project, experimental data are used to compose families of structural related molecules in order to build up a read across strategy. New QSAR models are built to be applied to dyes and describe several different end points.

The existing software in fact (EPISUITE, TOXTREE, OECHEM, DEREK, VEGA,…) are not applicable for dyes which are very large and complex substances, Ionic disconnected substances and actually outside of any existing Applicability Domain (AD).

A database on dyes containing all the structures, the existing data on physico-chemical properties and toxicities has been implemented.

Within this project, the whole set of substances is grouped based on their structural similarities and chemical reactivity.

In order to obtain reliable and predictive models several different multivariate techniques and chemometric methods are used. The structural description of the compounds is achieved using molecular descriptors of different classes calculated by means of main available software.

Each model is characterized by its Applicability Domain (AD), that is the chemical space where the model can be considered to provide reliable predictions.

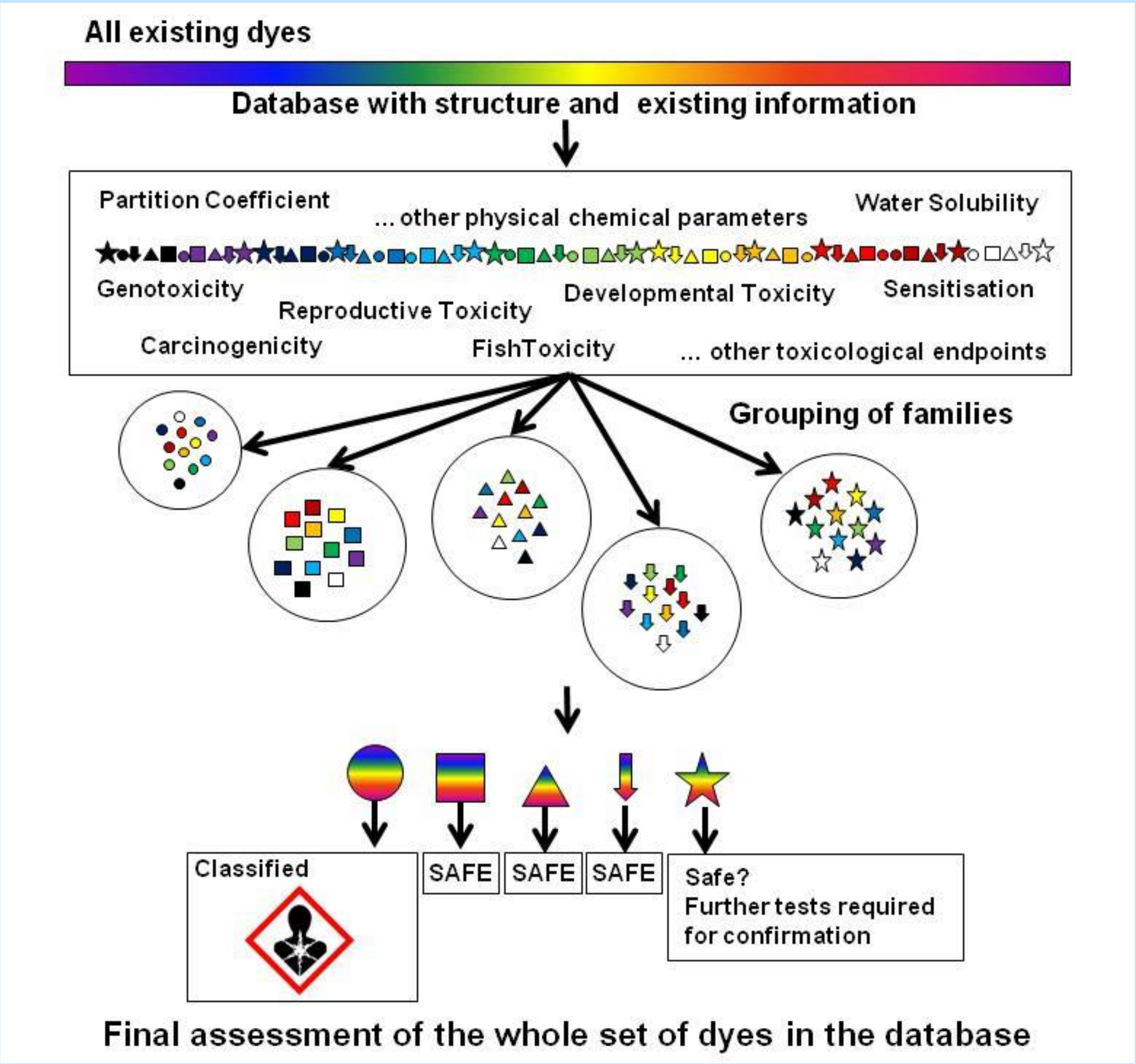
The strategy adopted to develop this set of QSAR models complies with the OECD principles for (Q)SAR validation, also implemented in REACH:

- A defined endpoint
- An unambiguous algorithm
- A defined domain of applicability
- Appropriate measures of goodness-of-fit, robustness and predictivity
- A mechanistic interpretation, if possible

A cross-validation procedure is carried out during the modeling stage in order to select the best subset of molecular descriptors and choose the best models on the basis of their predictive power.

Since the expected number of molecular descriptors that is calculated is very high, variable selection or variable reduction techniques is applied to obtain a suitable dataset. Once the best descriptors to represent dyes are identified, all the molecules are mapped on the chemical space defined by the selected molecular descriptors.

All the calculations will be performed by in-house statistical packages and, eventually, novel software specifically designed and implemented by MCQ group for this analysis.



Final assessment of the whole set of dyes in the database

## 5. SIEF MANAGER

SIEF Manager is the software that was developed by REACH&Colours Italia Srl to improve the exchange of information between the companies interested in REACH Registration of dyes.

It represents the interface between REACH IT system, the official web system used by the European Chemical Agency, (ECHA) and the activities of the REACH dyes consortia.

This software includes also the database of the existing and new performed studies including both public and proprietary data for more than 1000 dyes. Sharing of data was possible due to the common interest of many Companies in the REACH registration of such a large number of substances.

This database was interfaced with the QSAR modeling system.

Within two years the models for the following endpoints will be developed:

- Physical chemical properties
- Environmental fate (Biodegradability)
- Ecotoxycological Information (Daphnia, Algae and Fish)
- Toxycological information (Skin and Eye Irritation/corrosion)
- Skin sensitization
- Acute oral toxicity
- Mutagenicity

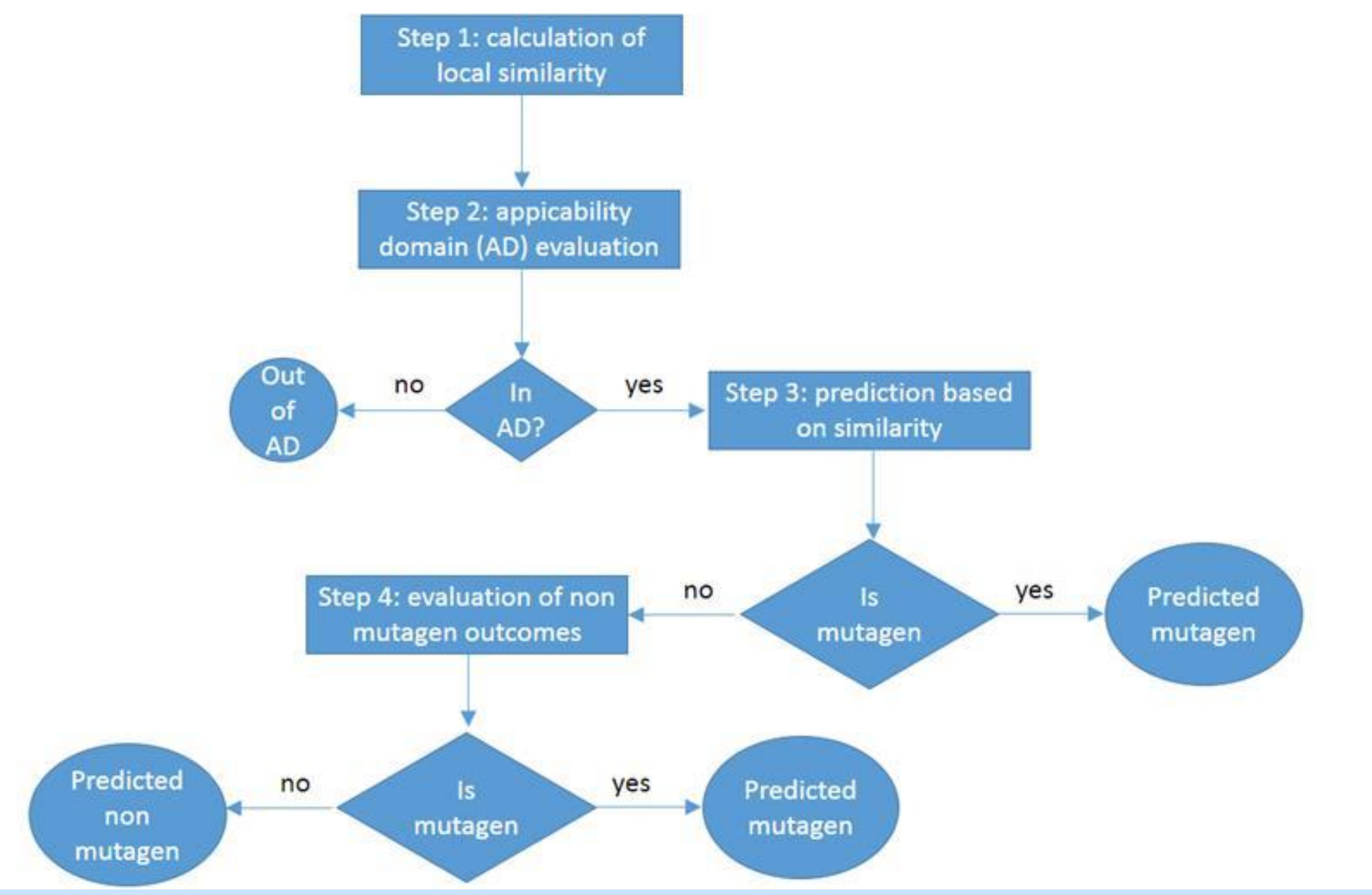
## 6. MODEL CLASSIFICATION PERFORMANCE

The model performance is evaluated on the basis of

- sensitivity, which is the percentage of mutagen molecules predicted as mutagen;
- specificity, which is the percentage of non mutagen molecules predicted as non mutagen;
- non-error rate (NER), which is the arithmetic mean of specificity and sensitivity;
- the percentage of not predicted molecules, which are the molecules with average JT similarity to the nearest molecules lower than the selected AD threshold (i.e., 0.3).

The confusion matrix is a specific table layout that allows visualization of the model classification performance of an algorithm.

Structure of the multi-step QSAR model – Example for MUTAGENICITY, Ames Test


Step 1: calculation of local similarity
Step 2: apicability domain (AD) evaluation
In AD?
no
Out of AD
yes
Step 3: prediction based on similarity
Is mutagen
yes
Predicted mutagen
no
Step 4: evaluation of non mutagen outcomes
Is mutagen
no
Predicted non mutagen
yes
Predicted mutagen

## 7. Confusion matrix derived for mutagenicity (Ames Test)

| | Predicted class | | Out of AD |
|---|---|---|---|
| | Mutagen | Non Mutagen | |
| Mutagen | 37 | 6 | 1 |
| Non Mutagen | 12 | 66 | 10 |

From the results the model gives the following quality indices:

sensitivity: 86%

specificity: 85%

NER: 85%

Out of AD: 8%

## 8. CONCLUSION

The developed model for Ames test, shows quite high classification parameters and the QSAR model can be considered able to discriminate mutagen and non-mutagen molecules without specific issues. Looking at misclassified molecules, the model looks well balanced, since sensitivity and specificity values are similar. These results demonstrate the validity of the QSAR approach which are promising for the next models which will be developed with the consequence of reduced number of new tests to be performed.

## 9. REFERENCE

Burden, F. R., Brereton, R. G., and Walsh, P. T. (1997). Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. *Analyst*, 122: 1015–1022.

Ferrari, T., and Gini, G. (2010). An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chemistry Central Journal* , 4, p. (Suppl 1)-S2

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model, 50: 742– 754.*