Review of methods to assess a QSAR Applicability Domain

Joanna Jaworska
Procter & Gamble
European Technical Center
Brussels, Belgium

and

Nina Nikolova – Jeliazkova
IPP
Bulgarian Academy of Sciences
Sofia, Bulgaria
Contents

• Why we need applicability domain?

• What is an applicability domain?
  – Training data set coverage vs predictive domain

• Methods for identification of training set coverage

• Methods for identification of predictive domain

• Practical use / software availability
Why we need applicability domain for a QSAR?

• Use of QSAR models for decision making increases
  – Cost & time effective
  – Animal alternatives

• Concerns related to quality evaluation of model predictions and prevention of model’s potential misuse.
  – Acceptance of a result = prediction from applicability domain

• Elements of the quality prediction
  – define whether the model is suitable to predict activity of a queried chemical
  – Assess uncertainty of a model’s result
QSAR models as a high consequence computing – can we learn from others?

• In the past QSAR research focused on analyses of experimental data & development of QSAR models
• The applicability domain QSAR definition has not been addressed in the past
  – Acceptance of a QSAR result was left to the discretion of an expert
  – It is no longer classic computational toxicology
  – currently the methods and software are not very well integrated with
• However, Computational physicists and engineers are working on the same topic
  – Reliability theory and Uncertainty analysis
  – increasingly dominated by Bayesian approaches
**What is an applicability domain?**

- Setubal report (2002) provided a philosophical definition of the applicability domain but not possible to compute code.

- Training data set from which a QSAR model is derived provides basis for the estimation of its applicability domain.

- The training set data, when projected in the model’s multivariate parameter space, defines space regions populated with data and empty ones.

- Populated regions define applicability domain of a model i.e. space the model is suitable to predict. This stems from the fact that generally, interpolation is more reliable than extrapolation.
Experience using QSAR training set domain as application domain

• **Interpolative predictive accuracy** defined as predictive accuracy within the training set is in general greater than **Extrapolative predictive accuracy**

  • The average prediction error outside application domain defined by the training set ranges is twice larger than the prediction error inside the domain.

• Note that it is true only **on average**, i.e. there are many individual compounds with low error outside of the domain, as well as individual compounds with high error inside the domain.

For more info see poster
What have we missed while defining applicability domain?

- The so far discussed approach to applicability domain addressed ONLY training data set coverage.
- Is applicability domain for 2 different models developed on the same data set same or different?
- Clearly we need to take into account model itself.
Applicability domain – evolved view

• Assessing if the prediction is from interpolation region representing training set does not tell anything about model accuracy
  - The only link to the model is by using model variables (descriptors)

• Model predictive error is eventually needed to make decision regarding acceptance of a result.
  - Model predictive error is related to experimental data variability, parameter uncertainty
  - Quantitative assessment of prediction error will allow for transparent decision making where different cutoff values of error acceptance can be used for different management applications
Applicability domain estimation – 2 step process

• Step 1 – Estimation of Application domain
  – Define training data set coverage by interpolation

• Step 2 – Model Uncertainty quantification
  – Calculate uncertainty of predictions, i.e. predictive error
Application domain of a QSAR

Training set of chemicals

Multivariate descriptor space

Legend:
- Estrogen binding training set
- HVPC database

Applicability domain = populated regions in multivariate descriptor space?
Application domain estimation

• Most of current QSAR models are not LFERs

• They are statistical models with varying degree of mechanistic interpretation usually developed *a posteriori*

• Statistical models application is confined to interpolation region of the data used to develop a model i.e. training set

• Mathematically, interpolation projection of the training set in the model’ descriptors space is equivalent to estimating a multivariate convex hull
Is classic definition of interpolation sufficient?

In reality often

- data are sparse and nonhomogenous;
  - Group contribution methods are especially vulnerable by the “Curse of dimensionality”
- Data in the training set are not chosen to follow experimental design because we are doing retrospective evaluations

Empty regions within the interpolation space may exist;

The relationship within the empty regions can differ from the derived model and we can not verify this without additional data;
Interpolation vs. Extrapolation

1D: parameter range determines interpolation region

>2D: is empty space within ranges interpolation?
**Interpolation vs. Extrapolation**

(Linear models)

Linear model: Predicted results within interpolation range do not exceed training set endpoint values.

Linear model - 2D – can exceed training set endpoint values even within ranges.
Approaches to determine interpolation regions

- Descriptor ranges
- Distances
- Geometric
- Probabilistic
Ranges of descriptors

- Very simple
- Will work for high dimensional models
  - Only practical solution for group contribution method
  - KOWIN model contains over 500 descriptors
- Cannot pick holes in the interpolated space
- Assumes homogenous distribution of the data
Distance approach

- Euclidean distance
  - Gaussian distribution of data
  - No correlation between descriptors

- Mahalanobis distance
  - Gaussian distribution of data
  - Correlation between descriptors
Probabilistic approach

• Does not assume standard distribution. Solution for general multivariate case by nonparametric distribution estimation

• The probability density is a most accurate approach to identify regions containing data

• Can find internal empty regions and differentiate between differing density regions

• Accounts for correlations, skewness
Bayesian Probabilistic Approach to Classification

• Estimate density of each data set
• Read off probability density value for the new point for each data set
• Classify the point to the data set with the highest probability value

Bayesian Classification Rule provides theoretically optimal decision boundaries with smallest classification error

• R.O.Duda and P.E.Hart. Pattern Classification and Scene Analisys, Wiley, 1973
Probability Density Estimation
multidimensional approximations

Assume $x_i$ independent
- Estimate 1D density by fast algorithm.
- Estimate nD density by product of 1D densities

Does not account for correlation between descriptors

• Extract principal components
  • Estimate 1D density by fast algorithm on each principal component
  • Estimate nD density by product of 1D densities

Accounts for linear correlations via PCA (rotation of the coordinate system)
Various approximations of Application domain may lead to different results

- (a) ranges
- (b) distance based
- (c) distribution based
Is it correct to say:

- “prediction result is always reliable for a point within the application region”?
- “prediction is always unreliable if the point is outside the application region”? NO!
Assessment of predictive error

- Assessment of the predictive error is related to model uncertainty quantification given the uncertainty of model parameters
  - Need to calculate uncertainty of model coefficients
  - Propagate this uncertainty through the model to assess prediction uncertainty
    - Analytical method of variances if the model is linear in parameters
      \[ y = ax_1 + bx_2 \]
    - Numerical Monte Carlo method
Methods to assess predictive error of the model

- Training set error
- Test error
- Predictive error
  - External validation error
  - Crossvalidation
  - Bootstrap
Conclusions

• Applicability domain is not a one step evaluation. It requires
  • estimation of application domain - data set coverage
  • Estimation of predictive error of the model

• Various methods exist for estimation of interpolated space, boundaries defined by different methods can be very different.

• Be honest and do not apply “easy” methods if the assumptions will be violated. It is important to differentiate between dense and empty regions in descriptor space, because Relationship within empty space can be different than the model and we can not verify this without additional data

• To avoid complexity of finding Application Domain after model development Use Experimental design before model development
Conclusions -2

• Different methods of uncertainty quantification exist, choice depends on the type of the model (linear, nonlinear)
Practical use/software availability

- For uncertainty propagation can we advertise Busy?
COVERAGE Application
Thank you!

Acknowledgements to Tom Aldenberg (RIVM)
Interpolation regions and applicability domain of a model

Example:

Two data sets, represented in two different 1D descriptors:

- green points
- red points

Two models (over two different descriptors $X_1$ and $X_2$).

- Linear model (green)
- Nonlinear model (red)

The magenta point is within coverage of both data sets.

Is the prediction reliable?

Coverage estimation should be used only as a warning, and not as a final decision of “model applicability”
Possible reasons for the error:

- Models missing important parameter
- Wrong type of model
- Non-unique nature of the descriptors

The models vs. The true relationship
Two data sets, represented in two different 1D descriptors:

- green points
- red points

Two models (over two different descriptors $X_1$ and $X_2$):

- Linear model (green)
- Nonlinear model (red)

The magenta point is OUT of coverage of both data sets.

Prediction could be correct, if the model is close to the TRUE RELATIONSHIP outside the training data set!