

ABSTRACT

In recent years, as the result of implementing the REACH Policy by the European Union, the attention of scientists and regulators has focused on establishing general validation principles for QSAR models in the context of chemical regulation (previously known as the Setubal principles, and more recently as OECD). This paper presents a brief analysis of some of the principles - unambiguous algorithm, Applicability Domain, robustness and predictivity. Some concerns related to QSAR algorithm reproducibility and an example of a fast check of the applicability domain for MLR models are given.

Common myths and misconceptions related to popular techniques for verifying internal predictivity, particularly for multilinear regression models (for instance cross-validation, bootstrap), are discussed with reference to statistical techniques commonly used for external validation. The differences in these two validating approaches have been highlighted, and evidence is presented that only models validated externally after internal validation can be considered reliable and applicable for both external prediction and regulatory purposes.

"Validation is one of those words...that is constantly used and seldom defined" (A. Feinstein)

From Setubal to OECD Principles

To facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information:

- | | |
|---|---|
| 1) be associated with a defined endpoint | 1) a defined endpoint |
| 2) take the form of an unambiguous and easily applicable algorithm; | 2) an unambiguous algorithm; |
| 3) ideally, have a mechanistic basis; | 3) a defined domain of applicability |
| 4) be accompanied by a definition of domain of applicability | 4) appropriate measures of goodness-of-fit, robustness and predictivity |
| 5) be associated with a measure of goodness-of fit (internal validation); | 5) a mechanistic interpretation, if possible; |
| 6) be assessed in terms of its predictive power by using data not used in the development of the model (external validation). | |

The two concepts of internal and external validation should be clearly separated [1-3]

Principle 2: an unambiguous algorithm

An algorithm must be, by definition, unambiguous: it must be clearly defined, exactly reproducible and continuously applicable also to new chemicals. BUT..... if the molecular descriptors are not reproducible?

Old Model [4] DRAGON Version

$\log\text{BCF} = -17.58 + 1.69 \text{ VIM.D.deg} - 0.45 \text{ nHAcc} + 15.65 \text{ MATS2m} - 0.36 \text{ GATS2e} - 1.64 (\pm 0.56) \text{ H6p}$ (1)
 n: 179 R2:0.79; Q2LOO:0.78; Q2LMO(50%):0.77; n: 59 Q2 EXT:0.88

Updated Model [5] DRAGON Version (some descriptors of Model 1 no more reproducible)

$\log\text{BCF} = -0.74 + 2.55 \text{ VIMD.deg} - 1.09 \text{ HIC} - 0.42 \text{ nHAcc} - 1.22 \text{ GATS1e} - 1.55 \text{ MATS1p}$ (2)
 n: 179 R2:0.81; Q2LOO:0.79; Q2BOOT:0.79; n:59 Q2EXT:0.88.

Polar narcotics Log (LC50) = -0.846 log Kow - 1.39.

Different values of logKow from different experimental methods or different software gave different results [6,7].

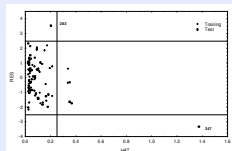
Which is the best, universally applied, log Kow value rendering this algorithm unambiguous?

Principle 3: a defined domain of applicability

The AD is a theoretical region in the space, defined by the descriptors of the model and the modeled response and thus by the nature of the chemicals in the training set, as represented, in each model, by the specific molecular descriptors.

Different approaches have been proposed [8], for the different typology of the models.

For an immediate and easy visualisation of the AD of MLR models the Williams plot: the plot of standardised cross-validated residuals (R) vs leverages (Hat diagonal) values (h) can be used.



A MLR model for polar narcotics in *Pimephales promelas* [6]:

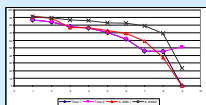
347 is wrongly predicted (> 3s, and high h leverage value): both a response outlier and out of model AD.
 Two other chemicals (squares at 0.35 h) are reliable as those of the near training chemicals.
 283 is wrongly predicted (>3s), but into the cut-off value of hat. It belongs to model AD: erroneous prediction due to a probable wrong experimental data

Principle 4: appropriate measures of goodness-of-fit, robustness and predictivity

In my report on Evaluation of different statistical approaches to the validation of QSAR [9]: some published models of Kulkarni et al [10] have been verified:

Ecotoxicity to *P. promelas* (LC50) of alcohols

n	Y	Model	Variable	Q ²	Q ² boot	Q ² ext	Q ² ext	Q ² ext
10	1	1	1	0.85	0.85	0.85	0.85	0.85
10	1	2	2	0.85	0.85	0.85	0.85	0.85
10	1	3	3	0.85	0.85	0.85	0.85	0.85
10	1	4	4	0.85	0.85	0.85	0.85	0.85



Ecotoxicity of aliphatics

Model	Variable	Q ²	Q ² boot	Q ² ext	Q ² ext
1	1	0.85	0.85	0.85	0.85
2	2	0.85	0.85	0.85	0.85
3	3	0.85	0.85	0.85	0.85
4	4	0.85	0.85	0.85	0.85

- overfitting and not predictive models
- CV can be too over-optimistic

MYTH: The Cross-Validation is sufficient to verify a model predictivity for NEW chemicals
 MISCONCEPTION: The above point is not exact, being CV an iterative process,

CV is NECESSARY but NOT SUFFICIENT

Baumann [2,3] pointed out that although the data used for validation are independent of the model-building process in each single split of the CV procedure, the resulting internal estimate of the prediction error is over-optimistic since the same data are repeatedly used to build and to assess the model [3,4].

The average of the performance measures of these parallel models, taken over several iterations, is considered the performance estimate of the final proposed model (the full model), whose equation is developed on all the chemicals (this is the unambiguous algorithm of Principle 2).

The parallel models, developed on reduced training sets for CV, are different models (in term of coefficients) from the final, full one.

In fact, some models with high internal predictivity, verified by internal CV methods (LOO, LMO, Bootstrap), (thus, applied to chemicals that iteratively contribute in the selection of molecular descriptors), can be externally less predictive or even absolutely unproductive [9], when applied on external chemicals, NEVER presented during the model development.

An example of this point is reported in the Table for the 16 models of a GA population of PAH mutagenicity models (TA100 on 48 chemicals: 31 in training and 17 in prediction set).

Model	R2	Q2	Q2boot	Q2ext
PW2 SCI1	0.57	0.54	0.26	0.27
Mo MATS2e	0.37	0.04	0.25	0.27
Mo GATS2e	0.37	0.05	0.25	0.27
Mo APW	0.25	0.08	0.21	0.21
Mo PW2	0.05	0.05	0.05	0.05
PW2 RC1	0.09	0.04	0.22	0.07
MO VED2	0.27	0.26	0.06	0.07
Mo LMO	0.26	0.19	0.21	0.24
BEAR HATS5a	0.63	0.61	0.27	0.29
PW2 IBC	0.14	0.09	0.16	0.02
VED2 Moa	0.27	0.26	0.06	0.07
HATS5a Moa	0.68	0.62	0.23	0
Mo MATS2m	0.25	0.23	0.04	0.02
ODD VED2	0.13	0.04	0.04	0.03
BEAR Moa	0.06	0.02	0.03	0.02

In CV the structural information of each chemical in the training set (the features represented by the selected descriptors) is taken into account in at least one validation run of the iterative process (none chemical is "new" at the end of this process).

For PREDICTIVITY → EXTERNAL validation, of internally validated models, on NEW chemicals is NECESSARY.

CONCLUSIONS

Principle 2: reproducibility of the descriptors and the application of the coefficients of the unambiguous algorithm of Principle 1 must be guaranteed.

Principle 3: a fast and simple way to verify the AD of a MLR model is the Williams plot, the plot of standardised cross-validated residuals vs leverages (derived from the Hat diagonal) values.

Principle 4: external validation on a significant and representative number of chemicals must always supplement internal validation for predictive QSAR models.

Concluding suggestion is again, after "The importance of being Earnest" [1]: externally validate, always and in a rigorous way, the internally stable models, in order to avoid the proposals of over-optimistic, erroneously called "predictive", QSAR models.

References

- [1] A. Tropsha, P. Gramatica, V.K. Gombar. *QSAR & Comb. Sci.*, 22, 69 (2003).
- [2] K. Baumann. *TrAC*, 22, 395 (2003).
- [3] K. Baumann, N. Stiefl. *J. Comput. Aid. Mol. Des.*, 18, 549 (2004).
- [4] P. Gramatica, E. Papa. *QSAR & Comb. Sci.*, 22, 374 (2003).
- [5] P. Gramatica, E. Papa. *QSAR & Comb. Sci.*, 24, 953 (2005).
- [6] E. Papa, F. Villa, P. Gramatica. *J. Chem. Inf. Model.*, 45, 1256 (2005).
- [7] E. Benfenati, et al. *Chemosphere*, 53, 1155 (2003).
- [8] T. I. Netzeva, et al. *ATLA*, 33, 155 (2005).
- [9] P. Gramatica (2004). <http://ecb.jrc.it/QSAR/> in QSARs/documents/public access
- [10] Kulkarni, S.A. et al. *SAR & QSAR in Environ. Res.*, 12, 565 (2001).