Validation and Predictivity of QSAR Models

Hugo Kubinyi
University of Heidelberg, c/o Donnersbergstrasse 9, D-67256 Weisenheim am Sand, Germany. E-Mail kubinyi@t-online.de

When QSAR started about 40 years ago [1], the quantitative description of structure-activity relationships (SARs) was in the foreground; prediction played only a minor role. A few physicochemical parameters, i.e. lipophilicity, expressed by log P or π values, electronic properties, expressed by σ, molar refractivity MR, steric properties, and/or parabolic lipophilicity terms were used in the correlations. Later, quantum chemical and geometrical parameters, connectivity values, electrotopological state parameters, WHIM parameters, and many others were tested (see, e.g. [2,3]), whether they are suited to explain SARs in a quantitative manner and whether the resulting models are capable to predict the activities of new analogs. As a consequence of so many (artificial) parameters and despite the fact that the use of too many parameters has been criticized already thirty years ago [4], the literature is now spoiled with, most probably, thousands of meaningless chance correlations.

Rules and conditions have been formulated to achieve valid correlations: meaningful variables shall be selected; the significance of the correlation and of each individual term in the regression model shall be justified by the appropriate statistical parameters (r, s, F values, sequential F test, and confidence intervals); the principle of parsimony shall be applied, i.e. results being more or less equal, the simplest model shall be chosen; not too many variables shall be tested and not too many variables shall be included in the final model [5]; some more detailed recommendations were published later [6,7]. In addition, crossvalidation [8], Y scrambling, and external (test set) predictivity are used as validation criteria.

In the past, the Selwood dataset [9] has become a standard in evaluating variable selection procedures (e.g. [10-12]); the biological activity data of 31 compounds are described by a few variables that are selected from 53 candidate variables. Whereas Selwood et al. and some later investigators (for a review, see [11]) were unable to derive the "best" models, evolutionary and genetic algorithms [10-12] uncovered models that may be regarded as the "best" ones, at least considering the common statistical parameters. The models with the highest F value, out of all possible 317,682 models with up to four X variables, is given by Eq. 1 [11,12]:

$$\begin{aligned}\log 1/C = &- 0.0000749\ (\pm 0.000030)\ \mathrm{MOFI\_Y} + 0.584\ (\pm 0.20)\ \mathrm{Log\ P} \\ &+ 1.514\ (\pm 0.91)\ \mathrm{Sum\_F} - 2.501\ (\pm 0.85) \qquad (1)\end{aligned}$$
$$(n = 31;\ r = 0.849;\ s = 0.460;\ F = 23.27;\ Q^2 = 0.647;\ s_{PRESS} = 0.518)$$

Eq. 1 corresponds to a very common situation in QSAR: a chemist synthesizes some 30 compounds; the biologist determines their activities; both ask a QSAR expert to derive a "good" model; the resulting model is justified by all statistical parameters, including confidence intervals of all regression coefficients and leave-one-out (LOO) crossvalidation. But the questions arise: is Eq. 1 really a valid model? Is it better than models derived from scrambled (or random) Y values, X values, or Y and X values? Are 53 variables too many to select from? Can our models predict a test set? Is there a relationship between internal (training set) and external (test set) predictivity?

For this purpose, systematic investigations of the Selwood dataset were performed. First, several thousand Y scrambling runs showed that less than 1% of all scrambled Y vectors have a correlation with the original Y values that is higher than $r^2 = 0.20$ (95% below $r^2 = 0.12$); correspondingly, a possible correlation between the original Y vector and the scrambled Y vectors can be neglected. Next, scrambled Y vectors were correlated with the variables of Eq. 1 (which is an inappropriate procedure; for every scrambled y vector, new X

variable combinations have to be tested); due to expectation, in 1,000 different runs only 1% of all models had F values >9.9 (95% with F <6.6; cf. Eq 1, F = 23.27; in these and all following calculations, F values are used as selection criterion). If a correct Y scrambling procedure is applied (i.e. Y scrambling first, then selection of the "best" model with up to four Y variables, from any of the 317,682 possible models), only 1% of all "best" models from 1,000 different Y scrambling runs had F values >16.0 (95% with F <12.3). Random (instead of scrambled) Y values, scrambled or random X values, and scrambled or random Y and X values gave corresponding results: in all cases (1,000 runs each, models selected from all 371,682 possible models), less than 1% of the models had F values >21.2 (95% with F <16.3; Figure 1; cf. Eq. 1, F = 23.27). Thus, neither scrambled nor random values produce a significant percentage of models that are as good as Eq. 1, indicating the significance of this QSAR model.

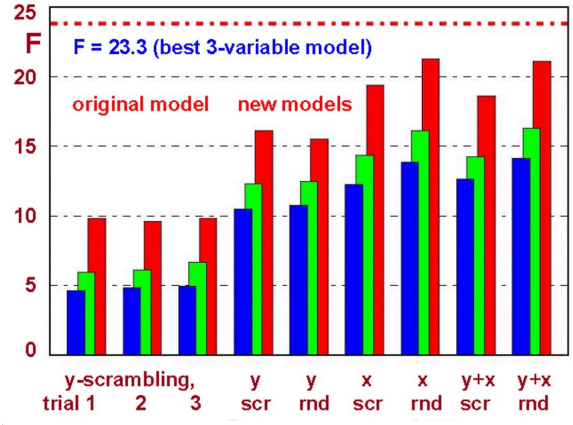

Figure 1. "Best" QSAR models with up to 4 variables for the Selwood dataset, using scrambled (scr) or random (rnd) Y, X, or Y and X values. The blue, green and red columns indicate the maximum F values for 90%, 95% and 99% of the models. In the first three cases, only the original X variables of Eq. 1 (upper dashed red line) were used, whereas in all other cases new "best" models were derived from all possible 317,682 models (1,000 runs for each case).

The next question is: are 53 variables a reasonable number to start from or are they too many? The X block randomization already shows that no "better" models are derived from such a number. However, there is a kind of linear increase of "good" results with an increasing number of random X variables: with 30 X variables to choose from, 95% of all models (again 1,000 runs, "best" models selected from any of the 317,682 possible models) have F values <13.0; with 40, 50, 60, 70 and 80 variables to select from, these F value limits are <13.8, <15.8, <16.3, <16.7, and <18.1. From this more or less linear increase one can extrapolate that only a selection from some 120-150 random variables may generate more than 5% "best" models with higher F values than Eq. 1.

So far, only fit is considered, i.e. a "reasonable" quantitative explanation of the underlying structure-activity relationship. But the real world is different: a chemist synthesizes some 20 compounds; the biologist determines the activities; both ask a QSAR expert to derive a "good" model. Can the model be used to predict the biological activity values of 10 compounds of the same dataset? For this purpose, random training and test set selections were performed; the size of the training sets varied from 30, 29, 28, 26, and 21 to 16 analogs (generating test sets of 1, 2, 3, 5, 10, and 15 analogs); in all cases, 1,000 random selections were performed and for every selection all 317,682 possible models were calculated to find the "best" model in each case. For the "real world" situation like the one described above, i.e. a training set of 21 compounds and a test set of 10 compounds, about 75% of all "best" models are justified by LOO crossvalidation (Q^2 >0.6; about 90% with Q^2 >0.5; Figure 2), indicating that in most cases of such a training/test set selection models can be obtained that may be considered to be significant from their internal predictivity.

The more important question is whether these models have external predictivity. Unfortunately, the answer is no, not at all. Only 1.6% of the "best" models produce r^2_{pred} values >0.6 (6.0% with r^2_{pred} >0.5) and only about 44% are better than average in their external predictivity (r^2_{pred} >0; Figure 3).

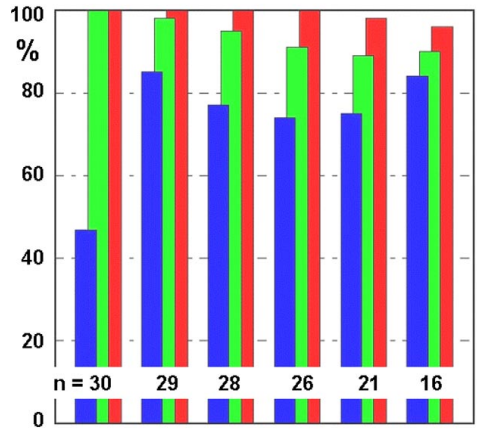

Figure 2: LOO crossvalidation results for the "best" QSAR models (out of 317,682 possible models with up to 4 variables; 1,000 runs in each case), starting from random training set selections with 30, 29, 28, 26, 21, or 16 compounds of the Selwood dataset. The blue, green and red columns indicate the percentage of models with better results than Q^2 = 0.6, 0.5, and 0.0.

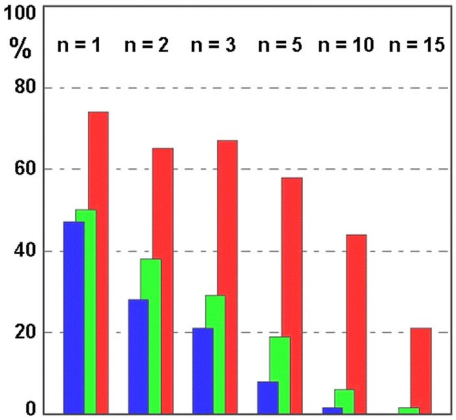

Figure 3: External predictivity of the "best" QSAR models of Figure 2 (n = 1, 2, 3, 5, 10, or 15 compounds in the test set). The blue, green and red columns indicate the percentage of models with better results than r^2_{pred} = 0.6, 0.5, and 0.0.

In the past, an inconsistency between internal and external predictivity has been observed in a few 3D QSAR and QSAR studies [13-15]. A first systematical investigation showed that, in general, there is no relationship between internal and external predictivity [16]: high internal predictivity may result in low external predictivity and vice versa. This effect, in the meantime being called the "Kubinyi paradox" [17,18], was also observed in other QSAR studies [19] as well as in a retrospective investigation of about 40 different 3D QSAR models [20]. For the Selwood dataset, Figure 4 shows that only 56 out of 1,000 "best" models have internal and external predictivity in a favorable range of Q^2 and r^2_{pred} = 0.6 - 1.0. All other models have either worse internal or worse external predictivity.

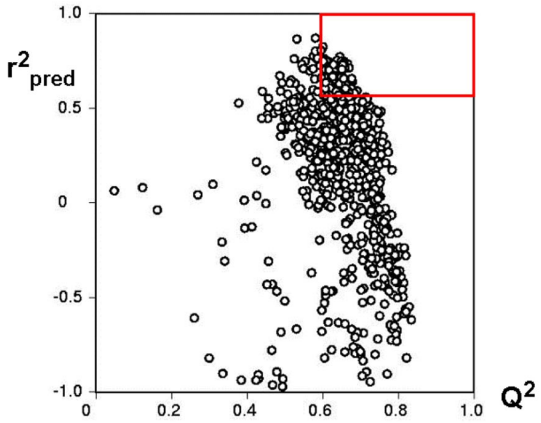

Figure 4: Comparison of internal and external predictivity of the "best" QSAR models, using a training set of 21 compounds and a test set of 10 compounds (cf. Figures 2 and 3); only 56 of 1,000 models fall into the favorable category Q^2 >0.6 and r^2_{pred} >0.6 (red box).

Neglecting for a moment the disappointing conclusion from this effect, it remains the question: what are the reasons and how can we derive better QSAR models? First, we should return to the recommendations of Topliss, Unger and Hansch [4,5], to include only reasonable variables, selected from small numbers, and to generate only models that have a sound biophysical background. But there remains another problem that can be demonstrated with another well-known dataset. Eq. 2 can be derived for the corticosteroid-binding globulin affinities of 31 steroids (4,5 >C=C< codes for a carbocyclic double bond between ring atoms 4 and 5) [15, 21, 22].

$$\log 1/CBG = 1.861\ (\pm 0.46)\ [4,5\ >C=C<] + 5.186\ (\pm 0.36) \qquad (2)$$

$$(n = 31;\ r = 0.838;\ s = 0.600;\ F = 68.28;\ Q^2 = 0.667;\ s_{PRESS} = 0.634)$$

As the dataset contains at least one outlier, compound # 31, it makes a significant difference whether this compound is included in the training set or in the test set. A training set selection of compounds # 1-21 (test set # 22-31) leads to Q^2 = 0.726 and r^2_{pred} = 0.477, with good internal and poor external predictivity; on the other hand, a training set selection of compounds # 1-12 and 23-31 (test set # 13-22) leads to Q^2 = 0.454 and r^2_{pred} = 0.909, with poor internal but excellent external predictivity [15].

As a conclusion, not only the recommendations of Topliss, Unger and Hansch should be followed, also the chemical space of training and test sets has to be analyzed; real outliers, with respect to congeneric character and structural similarity, have to be discovered and eliminated. Even then, prediction by QSAR models remains a risky procedure.

References

[1] Hansch C, Fujita T, ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc. 1964, 86: 1616-1626.

[2] Todeschini R, Consonni V, Handbook of Molecular Descriptors (Methods and Principles in Medicinal Chemistry, Volume 11, Mannhold R, Kubinyi H, Timmerman H, Eds.), Wiley-VCH, Weinheim, 2000.

[3] Todeschini R, Program DRAGON; www.disat.unimib.it/chm/Dragon.htm.

[4] Topliss JG, Costello RJ, Chance correlations in structure-activity studies using multiple regression analysis, J. Med. Chem. 1972, 15: 1066-1068.

[5] Unger SH, Hansch C, On model building in structure-activity relationships. A reexamination of adrenergic blocking activity of β-halo-β-arylalkylamines, J. Med. Chem. 1973, 16: 745-749.

[6] Wold S, Validation of QSAR's, Quant. Struct.-Act. Relat. 1991, 10: 191-193.

[7] Mager H, Mager PP, Validation of QSAR's: some reflections, Quant. Struct.-Act. Relat. 1992, 11: 518-521; Wold S, Answer to Mager and Mager, Quant. Struct.-Act. Relat. 1992, 11: 522.

[8] Cramer III RD, Bunce JD, Patterson DE, Frank IE, Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies, Quant. Struct.-Act. Relat. 1988, 7: 18-25; erratum 1988, 7: 91.

[9] Selwood DL, Livingstone DJ, Comley JCW et al, Structure-activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study, J. Med. Chem. 1990, 33: 136-142.

[10] Rogers D, Hopfinger AJ, Application of genetic function approximation (GFA) to quantitative structure-activity relationships, J. Chem. Inf. Comput. Sci. 1994, 34: 854-866.

[11] Kubinyi H, Variable selection in QSAR studies. I. An evolutionary algorithm, Quant. Struct.-Act. Relat. 1994, 13: 285-294.

[12] Kubinyi H, Variable selection in QSAR studies. II. A highly efficient combination of systematic search and evolution, Quant. Struct.-Act. Relat. 1994, 13: 393-401.

[13] Novellino E, Fattorusso C, Greco G, Use of comparative molecular field analysis and cluster analysis in series design, Pharm. Acta Helv. 1995, 70: 149-154.

[14] Norinder U, Single and domain variable selection in 3D QSAR applications, J. Chemom. 1996, 10: 95-105.

[15] Kubinyi H, A general view on similarity and QSAR studies, in: Computer-Assisted Lead Finding and Optimization, van de Waterbeemd H, Testa B, Folkers G, Eds.; VHChA and VCH, Basel, Weinheim, 1997; pp. 9-28

[16] Kubinyi H, Hamprecht FA, Mietzner T, Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices, J. Med. Chem. 1998, 41: 2553-2564.

[17] van Drie JH, Pharmacophore discovery - lessons learned, Curr. Pharm. Des. 2003, 9: 1649-1664.

[18] van Drie JH, Pharmacophore discovery: a critical review, in: Computational Medicinal Chemistry for Drug Discovery, Bultinck P, de Winter H, Langenaeker W, Tollenaere JP, Eds., Marcel Dekker, New York, 2004, pp. 437-460.

[19] Golbraikh A, Tropsha A, Beware of q^2, J. Mol. Graphics Modelling 2002, 20: 269-276.

[20] Doweyko A, 3D-QSAR illusions, J. Comput.-Aided Mol. Design, in print.

[21] Cramer III RD, Patterson DE, Bunce JD, Comparative molecular field analysis (CoMFA). I. Effect of shape on binding of steroids to carrier proteins, J. Am. Chem. Soc. 1988, 110: 5959-5967.

[22] Coats E, The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods, in: 3D QSAR in Drug Design. Recent Advances, Kubinyi H, Folkers G, Martin YC, Eds. Kluwer/ESCOM, Dordrecht, 1998, pp. 199-213; also published in Persp. Drug Discov. Design 1998, 12/13/14: 199-213.