Editorial

The Importance of Rigorous Statistical Practice in the Current Landscape of QSAR Modelling

> *The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved.*
>
> **Paul Dirac**

1. INTRODUCTION

The pharmacological, toxicological, and ecotoxicological effects of chemicals, *e.g*., drugs and xenobiotics on biological systems can be expressed by the following relation:

$$BR = f(S, B) \tag{1}$$

In the above equation, the biological response (BR) represents the normal biological effects produced as a result of exposure to the chemical, and B represents the relevant biochemical part of the target system which is perturbed by ligand to produce the measurable effect. It is widely believed that a major determinant of BR is the structure (S) of the ligand [1-3]. The structure becomes the sole determinant of the variation of BR from chemical to chemical when the biological system, B, is practically the same and there is alternation only in the structure of the ligand. Under such conditions Eq. (1) approximates to:

$$BR = f(S) \tag{2}$$

The prediction of property/bioactivity of molecules from their structure falls under the purview of the branch of science called quantitative structure-activity relationship (QSAR) [1-4].

The discovery of a life-saving drug currently needs US $400 million to 2 billion [5, 6]. Regulatory agencies like United States Environmental Protection Agency (USEPA) [7], have to routinely assess chemicals for their potential hazard to human and ecological health. Because testing of a large number of chemicals in the laboratory, both for drug design and environmental protection, is very expensive, properties calculated by QSARs are considered as acceptable alternatives to laboratory testing [8].

The field of QSAR had its humble beginning in 1868, when Crum-Brown and Fraser [9] observed that the constitution of permanently charged quaternary compounds determined their ‘physiological activity’. In the middle of the twentieth century, Hansch and coworkers [1] developed the linear free energy related (LFER) approach for QSAR using hydrophobicity, Hammett’s electronic parameter σ and various steric descriptors as independent variables for correlation. The LFER models are essentially property-property relationships (PPRs) where a set of physicochemical properties of chemicals is used to predict their biological/physical properties. Such methods worked well for the estimation of bioactivities of congeneric sets of chemicals. But in many cases, experimental physicochemical properties of many of the chemicals under investigation are not available. The PPR approaches are not very useful in such situations. A viable alternative to solve this quagmire is to use properties that can be computed directly from the structure of molecules without the input of any other experimental data. Topological, substructural, geometrical, and quantum chemical molecular descriptors belong to this group. For large sets of molecules, high level quantum chemical descriptors could be very demanding on computer resources. On the other hand, descriptors derived from topological aspects of chemical structures, *e.g*. topological indices and different types of substructures, have found wide applications in numerous QSAR studies. For a recent summary of the topic, please see the review by Basak [4].

2. MAJOR PILLARS OF QSAR

> *In God we trust; all others must bring data.*
>
> **W. Edwards Deming**

> *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.*
>
> **Ronald A. Fisher**

The four major pillars of the development of a good QSAR model model can be specified as:

a. Quality, accuracy and consistency of experimental data

b. Data on sufficient number of predictors and samples

c. Availability of relevant descriptors which quantify aspects of molecular structure relevant to the activity/ toxicity of interest

d. Use of appropriate methods for model building and validation.

Compared to about half a century ago, the landscape of availability and calculation of molecular descriptors is very different now, the main reason behind this being extensive utilization of high-performance computing and availability of relevant software for descriptor calculation and model building. Now available software like PaDEL [10], DRAGON [11], Molconn-Z [12], POLLY [13], and APProbe [14] can calculate a large number descriptors for a molecule very fast.

Theoretically calculated descriptors generally used in QSAR models can be put into four categories, ordered according to their computational complexity: (1) Topostructural (TS), (2) Topochemical (TC), (3) Three-Dimensional (3-D) and (4) Quantum Chemical (QC). 3-D and especially QC descriptors are computation-intensive, but several Hierarchical QSAR studies [4, 15] have found out that after TS and TC descriptors are utilized to build a QSAR model, adding 3-D and QC descriptors do marginal to no improvement to the predictive ability of the model.

There are some specific issues that need to be navigated when building a QSAR model. First, we must keep in mind that most of the descriptor-based data we work with are inherently high-dimensional (*i.e.* number of descriptors much larger than number of samples) and rank-deficient in nature. When we use a large number of calculated descriptors of a chemical to predict one of its physicochemical or biological property, it is only natural that the descriptors within a single class (*e.g.* all TS descriptors), or between two different classes (*e.g.* within TS and TC descriptors) collectively represent a substantial amount of overlapping information about the compound, and as a result are heavily correlated. In such scenarios standard statistical procedures often cease to be applicable. For example in linear regression, the probability distribution for the estimated vector of coefficients does not exist when there are more predictors than samples. In such situations, one should either explicitly try to select important variables before or during model building (*e.g.* forward selection, LASSO [16]), or attempt to find an optimal transformation that projects the data onto a lower-dimensional subspace before applying standard statistical methods (like principal component analysis (PCA), envelope methods [17]).

Since in QSAR analysis we are concerned about both interpretation of the model as well as its predictive ability, any model needs to be validated before being put to use. For a small number of samples, taking one compound out at a time, building model on remaining data, predicting activity of the held-out sample and repeating for all compounds is computationally tractable. This is the so-called Leave-one-out (LOO) cross-validation. As sample size grows, the *k*-fold cross-validation is more economical instead: divide the samples into *k* equal-sized parts and take as holdout each part to obtain predictions. Compared to a cross-validation approach, external validation, *i.e.* choosing one single train-test set split based on some criterion or domain knowledge suffers from drawbacks like leaving out information from held-out compounds, lack of any theoretically sound method of determining similarity between training and test sets and inability to control for any inadvertent human bias. In fact, Hawkins *et al.* show that unless the dataset to be modeled contains a large number of samples, cross-validation is theoretically superior to external validation in estimating the true predictive ability of a QSAR model [18].

When doing cross-validation using models that involve one or more tuning parameters, or involve predictor selection/ dimension reduction, one also needs to keep in mind to repeat these pre-model formulation steps for each split of the data. If the full data is used to tune the model or select important predictors, and these options are kept constant in the subsequent cross-validation, it results in overestimation of the predictive power of the QSAR model. Hawkins *et al.* named the cross-validated q^2 obtained from such procedures as naïve q^2 [19, 20]. This happens because in the first step, information from the holdout sample/ split is used in determining the parameters, which are later plugged in the model to predict activity in that very compound or subset of compounds. Compared to this procedure, the true q^2 obtained from a two-stage procedure (or 'two-deep' cross validation) often requires significantly larger computation time, but due to the availability of considerable computational resources to present-day researchers, this is not too much of a problem.

Finally in the context of QSAR, one needs to define the domain of applicability of a developed model, and model only those new compounds using it whose predictor information falls within this domain. One can either explicitly define the interpolation region in the predictor space by applying some summarization method on the training set of predictors, like bounding box, principal component analysis, convex hulls *etc*., or take a distance-based approach to find out the relative similarity of a new compound to the training data, and only fit the predictive model on compounds that are similar enough. A rigorous overview and application of the above concept can be found in Sahigara *et al.* [21].

3. CONCLUDING REMARKS

The criteria discussed above will enable researchers to develop effective QSARs using computed chemodescriptors. In the post-genomic era, propelled by the Human Genome Project, the omics sciences, *viz*., genomics, proteomics, metabolomics, are generating biological attributes or biodescriptors for many chemicals. One of us (SCB) has been involved for over a decade in the use of biodescriptors as well as combined set of chemo- and biodescriptors in predicting pharmacological and toxicological activities of chemicals using robust statistical methods [3]. We sincerely hope the scientific community will be benefited by following the aforementioned procedures regarding QSAR formulation and validation.

REFERENCES

[1] Hansch, C.; Leo, A. *Exploring QSARs: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society: Washington, D.C., **1995**.

[2] Kier, L.B.; Hall, L.H. *Molecular Structure Description: The Electrotopological State*, Academic Press: San Diego, CA, **1999**.

[3] Basak, S.C. Role of mathematical chemodescriptors and proteomics-based biodescriptors in drug discovery. *Drug Dev. Res.*, **2010**, *72*, 1-9.

[4] Basak, S.C. Mathematical descriptors for the prediction of property, bioactivity, and toxicity of chemicals from their structure: a chemical-cum-biochemical approach. *Curr. Comput. Aided Drug Des.*, **2013**, *9*, 449-462.

[5] Adams, C.; Brantner, V. Estimating the cost of new drug development: is it really $802 million? *Health Aff. (Millwood)*, **2006**, *25*, 420-428.

[6] DiMasi, J.; Hansen, R.; Grabowski, H. The price of innovation: new estimates of drug development costs. *J. Health Eco.*, **2003**, *22*, 151-185.

[7] United States Environmental Protection Agency; http://www.epa.gov/nrmrl/std/cppb/qsar/index.html

[8] Benigni, R.; Bossa, C.; Tcheremenskaia, O.; Giuliani, A. Alternatives to the carcinogenicity bioassay: *in silico* methods, and the *in vitro* and *in vivo* mutagenicity assays. *Expert Opin. Drug Metab. Toxicol.*, **2010**, *6*, 1-11.

[9] Crum-Brown, A.; Fraser, T.R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. Roy. Soc. Edinburgh*, **1868**, *25*, 151-203.

[10] Yap, C.W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **2011**, *32*(7), 1466-1474.

[11] Todeschini, R.; Consonni, V.; Mauri, A.; Pavan M. *DRAGON - Software for the Calculation of Molecular Descriptors, Version 5.4*, Talete srl, Milan, Italy, **2006**.

[12] Hall Associates Consulting, *Molconn-Z Version 4.05*, Quincy, MA, **2003**.

[13] Basak, S.C.; Harriss, D.K. Magnuson, V.R. *POLLY v. 2.3*, Copyright of the University of Minnesota, **1988**.

[14] Basak, S.C.; Grunwald, G.D. *APProbe*, Copyright of the University of Minnesota, **1993**.

[15] Basak, S.C.; Majumdar, S. Prediction of mutagenicity of chemicals from their calculated molecular descriptors: a case study with structurally homogeneous *versus* diverse datasets. *Curr. Comput. Aided Drug Des.*, **2015**, *in press*.

[16] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning -Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, NY, **2008**.

[17] Cook, R.D.; Li, B.; Chiaromonte, F. Envelope models for parsimonious and efficient multivariate linear regression. *Stat. Sinica*, **2010**, *20*, 927-1010.

[18] Hawkins, D.M.; Basak, S.C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.*, **2003**, *3*, 579-586.

[19] Hawkins, D.M.; Basak, S.C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 37-44.

[20] Natarajan, R.; Basak, S.C.; Mills D.; Kraker, J.J.; Hawkins, D.M. Quantitative structure-activity relationship modeling of mosquito repellents using calculated descriptors. *Croat. Chem. Acta*, **2008**, *81*(2), 333-340.

[21] Sahigara, F.; Mansouri, K.; Ballabio, D; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, **2012**, *17*, 4791-4810.

Subhash C. Basak
(***Editor-in-Chief***)
Natural Resources Research Institute and
Department of Chemistry & Biochemistry
University of Minnesota Duluth
Duluth
MN 55811
USA
E-mail: sbasak@nrri.umn.edu

Subhabrata Majumdar
School of Statistics
University of Minnesota Twin Cities
Minneapolis
MN 55414
USA
E-mail: majum010@umn.edu