Pure & Appl. Chem., Vol. 69, No. 2, pp. 245–248, 1997.
Printed in Great Britain.

QSPR as a means of predicting and understanding chemical and physical properties in terms of structure

Alan R. Katritzky[a], Mati Karelson[b] and Victor S. Lobanov[a]

[a]*Florida Center of Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611-7200, USA,* [b]*Department of Chemistry, University of Tartu, 2 Jakobi St., Tartu EE2400, Estonia*

Biography: Alan R. Katritzky (b. 1928, London, U. K.) is Kenan Professor of Chemistry and Director of the Institute for Heterocyclic Compounds at the University of Florida. A light-hearted account of his life appeared in *J. Het. Chem*, **31**, 569-602 (1994), and an overview of his scientific work in *Heterocycles*, **37**, 3-130 (1994). He is actively engaged in research and teaching, editing, industrial consulting, international travel, and windsurfing.

Victor S. Lobanov, born in 1966 in Yekaterinburg, Russia, received his M. Sc. in chemistry at the Moscow State University, Russia in 1988 and his Ph.D. in computational chemistry from Tartu University, Estonia in 1995. He is currently a postdoctoral fellow at University of Florida with Professor Katritzky.

Mati Karelson (b. 1948) is the Professor and Head of Theoretical Chemistry at the University of Tartu, Estonia. He received his Ph. D. in physical organic chemistry in 1975. His research has dealt with the theory of solvent effects, the foundations of QSAR/QSPR, and the development of the respective computer software. He is a member of the International Society of Quantum Biology and Pharmacology and the New York Academy of Sciences. In 1994, he was nominated as a Courtesy Professor in Chemistry at the University of Florida.

It is a fundamental tenet of chemistry that the structural formula of any compound contains coded within it all that compound's chemical, physical, and biological properties. Physical organic chemistry in the 21st century will, we believe, become increasingly oriented towards elucidating in detail how these properties are determined by the structure, such prior considerations thus enabling subsequent experimentation to be concentrated in the most promising directions. It is clear that for many reasons, including the physical limitations of computer technology and the absence of a proper theoretical basis, quantitative calculations of chemical and physical properties of chemical compounds from first principles will not be achievable in the foreseeable future. Therefore, the development of alternative approaches to find quantitative mathematical relationships between the intrinsic molecular structure and observable properties of chemical compounds will be of increasing importance in the chemistry of the 21st century.

Classical physical organic chemistry has long been concerned with the correlation of chemical properties in terms of structure. However, most such work in the 20th century has been carried out with co-generic sets of compounds in which just one structural feature is changing at any one time. Numerous linear free energy relationships starting from those of Hammett thus resulted and have given considerable insight into organic chemical mechanisms. We believe that in the 21st century, the quantitative structure-property relationship (QSPR) approach will become the tool of choice for many academic and industrial chemists.

In the area of biological properties, the quantitative structure-activity relationship (QSAR) methodology has already become an essential tool in all serious medicinal chemistry. All major pharmaceutical companies have considerable effort directed towards elucidating the effect of structure on particular biological properties, particularly of medicinally active compounds. By contrast, the application of similar QSPR techniques to the elucidation of the ways in which structure determines chemical and physical properties has been less developed. In this area, the ADAPT program of Jurs (1) and the SPARC software developed at the U. S. Environmental Protection Agency (2) were visionary and led to significant advances in different areas of

QSPR and QSAR. More recently, the CODESSA program has been described (3, 4) and used in a variety of situations. The CODESSA program combines a large variety of classical non-empirical molecular descriptors (a non-empirical descriptor is any numerical quantity that can be derived solely from the structure of a compound) together with more novel quantum chemical and combined descriptors and invokes both standard and advanced statistical data treatment techniques such as multiple and nonlinear regression, factor analysis, and heuristic methods for the development of QSPR correlations in very large descriptor spaces. Importantly, all molecular descriptors used in this software are derived solely from the molecular structure, without requiring any experimental information. Therefore, an enormous attraction of QSPR is that it potentially combines the ability to predict chemical and physical properties of as yet unmeasured or unknown compounds with the ability to understand just how the structure influences a particular chemical and physical property.

The first results on the QSPR development using the CODESSA approach have been very encouraging. Thus, a correlation of gas chromatographic response factors and retention times, using a set of 152 diverse organic compounds with a wide range of functional groups, gave good correlations (see Figs. 1 and 2) with a very limited number of theoretical molecular descriptors (5). Perhaps more importantly, the descriptors found to be significant are physically meaningful. Thus, for retention times, the most important parameters are the α-polarizability and the minimum valency at an H atom, reflecting the intermolecular induction and dispersion interactions, and hydrogen bonding between the injected compound and the medium of the gas chromatographic column, respectively. For response factors, the most important descriptor is the relative weight of "effective" carbon atoms in the compound, which was defined from theoretical considerations of the decomposition of the chemical compound within the flame ionization detector.

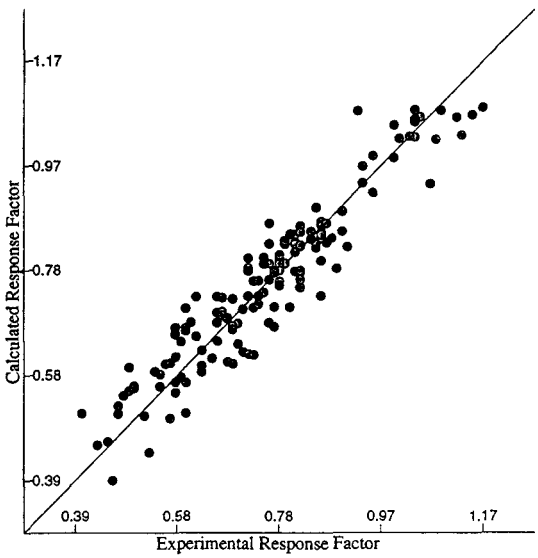

Fig. 1. The calculated vs. experimental gas chromatographic response factors for 152 diverse chemical compounds using a six parameter QSPR correlation with theoretical molecular descriptors (5) (R^2 = 0.892, F = 200, s = 0.054)

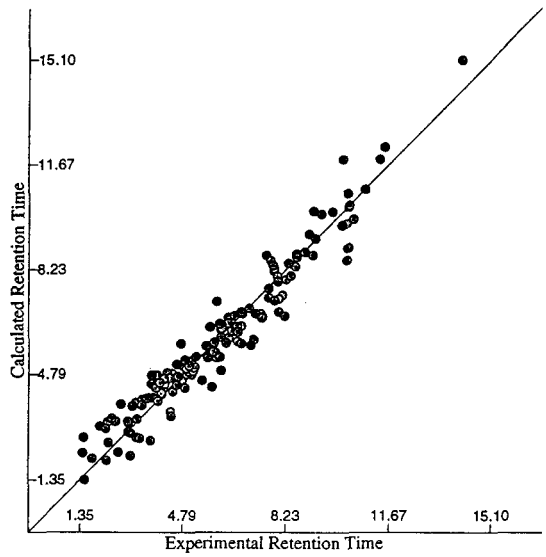

Fig. 2. The calculated vs. experimental gas chromatographic retention times for 152 diverse chemical compounds using a six parameter QSPR correlation with theoretical molecular descriptors (5) (R^2 = 0959, F = 362, s = 0.52)

Investigation of the boiling points of 298 diverse organic compounds found a two-parameter model with R^2 of 0.953, corresponding to an average predicted error of 3.0% (6). By adding two more parameters, the model could be improved to R^2 of 0.972 and a predicted error of 2.3%, which is very close to the average experimental error of 2.1% (*cf.* also Fig. 3). Obviously, the first two descriptors are the most important and they are highly physically significant, being the cubic root of the gravitation index and the hydrogen bonding donor charged surface area. One of them describes the effective dispersion interaction (the corresponding one-parameter correlation of hydrocarbon boiling points is excellent, with an R^2 of 0.965) whereas the second descriptor accounts for the hydrogen bonding in molecular liquids. Most importantly, the two descriptor

model developed solely on the basis of organic compounds offers excellent predictions for simple inorganic molecules. For example, the predicted boiling point for ammonia is 267 K (exptl. 240 K), for hydrogen fluoride 296 K (exptl. 293 K), and, most interestingly, for water 371 K (exptl. 373.15 K). We now have a clear physical insight into just how the structure affects the boiling point. In order to get further accuracy, it is necessary to use a larger data set and to correlate the residuals from the base correlation to find further less-obvious structural influences. Thus, in addition to considerable potential benefit to chemical engineers, we believe we can also use this QSPR technique to provide clear physical insight.

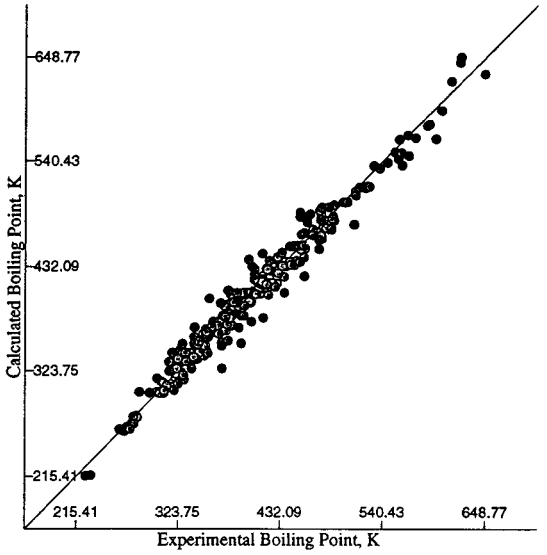

Fig. 3. The calculated vs. experimental boiling points for 298 organic compounds of diverse chemical structure using a four parameter QSPR correlation with theoretical molecular descriptors (6) (R^2 = 0.972, F = 2566, s = 12.4 K)

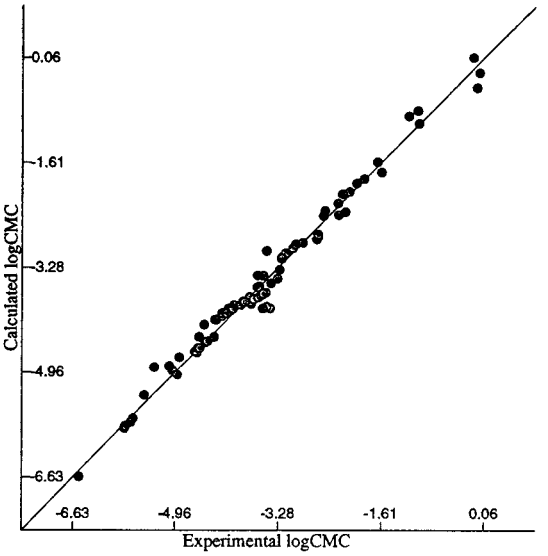

Fig. 4. The calculated vs. experimental critical micelle concentrations (CMC) for 77 nonionic surfactants using a three parameter QSPR correlation with theoretical molecular descriptors (7) (R^2 = 0.983, F = 1433, s = 0.031)

Another completely different area of high technological importance where this has been accomplished is in the area of surfactant science. The properties of 77 diverse nonionic surfactants are very well correlated (R^2 = 0.983) with only three theoretical molecular fragment descriptors (7) (*cf.* Fig. 4). Significantly, two of these descriptors are topological expressions for the hydrophobic tail and the third descriptor represents the size and polarity of the hydrophilic head. An immediate conclusion is that with the very high R^2 value, the correlation can be used confidently for the prediction of the CMC for unknown or unmeasured surfactants. Furthermore, the design of new surfactants is immeasurably helped by the understanding of the topological modes by which the hydrophobic fragment influences the CMC.

CODESSA has also been applied to many other chemical and physical properties of compounds. A large part of this work is proprietary but some of the results have been released. For instance, the melting points, boiling points, flash points, gas-chromatographic retention indices, and octanol-water partition coefficients of extended sets of substituted pyridines were successfully correlated with the theoretical molecular descriptors (8). In each case, the most important descriptors involved in the correlation have definite physical meaning and are clearly connected with the particular property studied.

A very significant extension of the QSPR work has been its application to the prediction of the properties of polymers (9). The non-empirical technique used here is the calculation of descriptors in the usual way for several lower oligomers for each polymer and then extrapolation to give the descriptor values for molecules of polymer size. This approach was used successfully to predict glass transition temperatures for a set of low molecular weight polymers and copolymers.

As scientists, today we are increasingly urged to do science of relevance to society. QSPR allows us to carry out science that will undoubtedly help us to make the production of new molecules useful in all facets of life and society more cheaply, more efficiently, and in a more environmentally friendly manner. At the same time, it offers the highest intellectual challenges for the development of meaningful relationships, novel theories, and deeper understanding into the molecular nature of the world in which we live.

REFERENCES

1. A. J. Stuper, W. E. Brugger, and P. C. Jurs, *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience, New York (1979).
2. S. H. Hilal, L. A. Carreira and S. W. Karickhoff. *Theor. Comput. Chem.* **1**, 291 (1994).
3. A. R. Katritzky, V. S. Lobanov and M. Karelson. *Chem. Soc. Rev.*, **24**, 279 (1995).
4. A. R. Katritzky, V. S. Lobanov and M. Karelson. *CODESSA Training Manual* (1995).
5. A. R. Katritzky, E. S. Ignatchenko, R. A. Barcock, V. S. Lobanov and M. Karelson. *Anal. Chem.* **66**, 1799 (1994).
6. A. R. Katritzky, L. Mu, V. S. Lobanov and M. Karelson. *J. Phys. Chem.*, **100**, 10400 (1996)..
7. P. T. Huibers, V. S. Lobanov, A. R. Katritzky, D. O. Shah and M. Karelson. *Langmuir* **12**, 1462 (1996).
8. R. Murugan, M. P. Grendze, J. E. Toomey, Jr., A. R. Katritzky, M. Karelson, V. S. Lobanov and P. Rachwal. *CHEMTECH* **24**, 17 (1994).
9. A. R. Katritzky, P. Rachwal, K. W. Law, M. Karelson and V. S. Lobanov. *J. Chem. Inf. Comp. Sci.*, **36**, 879 (1996).