MOLECULAR DESCRIPTORS USED IN QSAR continued

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches, as well as in quality control, being the way molecules, thought of as real bodies, are transformed into numbers, allowing some mathematical treatment of the chemical information contained in the molecule. This was defined by Todeschini and Consonni as:

"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."

By this definition, the molecular descriptors are divided into two main categories: **experimental measurements**, such as log P, molar refractivity, dipole moment, polarizability, and, in general, physico-chemical properties, and **theoretical molecular descriptors**, which are derived from a symbolic representation of the molecule and can be further classified according to the different types of molecular representation.

The main classes of theoretical molecular descriptors are: 1) **0D-descriptors** (i.e. constitutional descriptors, count descriptors), **1D-descriptors** (i.e. list of structural fragments, fingerprints), **2D-descriptors** (i.e. graph invariants), **3D-descriptors** (such as, for example, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, quantum-chemical descriptors, size, steric, surface and volume descriptors), **4D-descriptors** (such as those derived from GRID or CoMFA methods, Volsurf).

Invariance properties of molecular descriptors

The invariance properties of molecular descriptors can be defined as the ability of the algorithm for their calculation to give a descriptor value that is independent of the particular characteristics of the molecular representation, such as atom numbering or labeling, spatial reference frame, molecular conformations, etc. Invariance to molecular numbering or labeling is assumed as a minimal basic requirement for any descriptor.

Two other important invariance properties, translational invariance and rotational invariance, are the invariance of a descriptor value to any translation or rotation of the molecules in the chosen reference frame. These last invariance properties are required for the 3D-descriptors.

Degeneracy of molecular descriptors

This property refers to the ability of a descriptor to avoid equal values for different molecules. In this sense, descriptors can show no degeneracy at all, low, intermediate, or high degeneracy. For example, the number of molecule atoms and the molecular weights are high degeneracy descriptors, while, usually, 3D-descriptors show low or no degeneracy at all.

Basic requirements for optimal descriptors

1 Should have structural interpretation

2 Should have good correlation with at least one property

3 Should preferably discriminate among isomers

4 Should be possible to apply to local structure

5 Should possible to generalize to "higher" descriptors

6 Should be simple

7 Should not be based on experimental properties

8 Should not be trivially related to other descriptors

9 Should be possible to construct efficiently

10 Should use familiar structural concepts

11 Should change gradually with gradual change in structures

12 Should have the correct size dependence, if related to the molecule size

TRANSPORT THROUGH AN ORGANISM

The question arises as to why activity first increases and then decreases as hydrophobicity increases. Put simply, a chemical with a low partition coefficient will move only slowly from water to lipid, and therefore will arrive slowly at a receptor site after its random walk through a number of lipid membranes. Conversely, a chemical with a high partition coefficient will move quickly from water to lipid, but will leave lipid for the next aqueous phase slowly, therefore again arriving slowly at the receptor site. So the chemical with the most rapid movement will be one with intermediate values of both water→lipid and lipid→water partition coefficient. There are a number of other possible reasons for non-rectilinearity, such as different distribution in different compartments of an organism, steric restrictions at a receptor site, and increased metabolism of more hydrophobic xenobiotics as the organism seeks to excrete them

Penniston et al. (1969) devised a computer-based model of partition-based transport through 20 alternating aqueous and lipid compartments to demonstrate this. Partition coefficient is an equilibrium constant, and can be defined as the ratio of the rate constant of transfer from water to lipid and that of transfer from lipid to water:

$$P = k_{wo}/k_{ow} \qquad (16)$$

By setting up differential equations for each compartment, and following the "administration" of a single dose into the first compartment at time = 0, they were able to show that after a reasonable time, the concentration in the nth compartment rose as P increased, and then fell again, exactly as had been observed for the activity in biological systems.

McFarland (1970) developed a similar model, which was modified by Kubinyi (1977), who developed a bilinear model of activity versus log P rather than the parabolic model proposed by Hansch and Fujita (1964). Kubinyi's bilinear equation is:

$$\log 1/C = a\log P - b(\beta P + 1) + c \quad (17)$$

where a, b, β and c are constants. Kubinyi reported numerous examples where his bilinear model gave better fits than did the parabolic model; for example, for the narcotic action of 10 alcohols on the frog ventricle, the bilinear model yielded $R^2 = 0.996$, $s = 0.133$, whilst the parabolic model gave $R^2 = 0.956$, $s = 0.458$. The value of the Kubinyi equation is that it can accommodate different slopes of the upward and downward legs of a biphasic curve; a disadvantage is that it is not simple to use, as the β value has to be obtained by iteration.

Dearden and Townend (1978) also extended the Penniston model, and showed that both duration of action and time to maximal response varied parabolically with hydrophobicity, in each case going through a minimum at some intermediate value of log P. Also, although the Penniston model produced a parabolic type curve of activity against hydrophobicity at a given time after dosage, as had been found to exist in practice (Hansch & Fujita, 1964), Dearden and Townend (1983) showed that, at time to maximal response, the plot of activity against hydrophobicity initially rose, then levelled out instead of falling. This suggests that hydrophobic drugs could be significantly more potent than is indicated by measurement of potency at a fixed time after dosage.

QSAR IN ENVIRONMENTAL SCIENCES

As shown above, much of the very early work on the relationship between chemical structure and biological activity involved toxicity rather than pharmacological activity. It is therefore somewhat surprising that, following the key paper by Hansch et al. (1962), it was in pharmacology that QSAR was most widely used. In fact, it was not until the early 1980s that QSAR began to be used to any degree in environmental sciences.

One of the earliest environmental QSAR studies (Neely, Branson, & Blau, 1974) found a good rectilinear correlation between bioconcentration factor (BCF) and partition coefficient for some organic environmental pollutants:

$$\log BCF = 0.124 + 0.542 \log P \quad (18)$$

$$n = 8 \; r^2 = 0.899 \; s = 0.342$$

where BCF = ratio of steady-state concentration of chemical in trout muscle to that in water. Hence BCF can be regarded as an organism-water partition coefficient.

An important paper by Könemann (1981) showed that fish toxicity could be correlated well with log P for industrial chemicals acting via non-polar narcosis:

$$\log 1/LC_{50} = -4.87 + 0.871 \log P \quad (19)$$

$n = 50\ r^2 = 0.976\ s = 0.237$

where LC_{50} = concentration to kill 50% of guppies in a given time.

The log P values of the 50 chemicals ranged from –1.35 to 5.69, and Könemann commented that he was unable to measure LC_{50} values of chemicals with log P > 6, presumably because of low aqueous solubility.

A significant result of Könemann's work is the excellent straight-line correlation that he obtained, with no sign of the curvature often found with single-dosage tests on animals. The two situations are quite different, with the fish being supplied with a constant "dose" of the test chemical in the surrounding aqueous milieu. Dearden and Townend (unpublished information) found that their computer-based model of distribution of xenobiotics in an organism (Dearden & Townend, 1978) generated the same type of correlation, namely a straight-line correlation with log P when a constant concentration was maintained in compartment 1 of the model.

Further work has shown (Schultz, Lin, & Arnold, 1991; Cronin, 2003) that different mechanisms of action required different QSAR models. Using the toxicities of chemicals to the aquatic ciliate *Tetrahymena pyriformis*, they reported the following QSARs:

Non-polar narcosis: $\log 1/IGC_{50} = 0.74 \log P - 1.86$ (20)

$n = 148\ r^2 = 0.96\ s = 0.21\ F = 3341$

Polar narcosis: $\log 1/IGC_{50} = 0.59 \log P - 0.94$ (21)

$n = 119\ r^2 = 0.87\ s = 0.26\ F = 806$

Uncouplers: $\log 1/IGC_{50} = 0.40 \log P - 0.19$ (22)

$n = 12\ r^2 = 0.82\ s = 0.25\ F = 52$

Electrophiles: $\log 1/IGC_{50} = 0.60 \log P - 0.33\ E_{LUMO} - 1.00$ (23)

$n = 239\ r^2 = 0.80\ s = 0.34\ F = 476$

where E_{LUMO} = energy of the lowest unoccupied molecular orbital.

It should be noted that the rectilinear correlations of aquatic toxicity with log P are obtained when the system has reached steady state conditions; typically an aquatic toxicity test is conducted over several days – a 96-hour test period is often used.

The correlations of equations similar to those of equations 20 – 22 converge at log P values of 5-6 (Dearden, 2002).

Following increasing interest in environmental QSAR in the 1980s, a series of international workshops on QSAR in the environmental sciences was started in 1983 by Klaus Kaiser. As with the European QSAR symposia, they were initially held triennially, but from 1988 have been held biennially. Table 3 gives details of those held to date, and the publications arising from them. From 2004 their scope was expanded to include health sciences.

The predictive ability of QSAR models, particularly in environmental fields, opened up the possibility of predicted toxicities and physicochemical properties being used in regulatory matters. In 2002 a meeting of QSAR experts was held in Setúbal, Portugal, to produce a set of QSAR guidelines that would ensure that QSAR predictions used for regulatory purposes were properly validated. Those guidelines were then taken under the aegis of the Organisation for Economic Co-operation and Development (OECD, 2007). They are:

To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. *A defined endpoint;*
2. *An unambiguous algorithm;*
3. *A defined domain of applicability;*
4. *Appropriate measures of goodness-of-fit, robustness and predictivity;*
5. *A mechanistic interpretation, if possible.*

These criteria have assumed more importance with the implementation of the REACH (Registration, Evaluation, Authorisation, and Restriction of Chemicals) legislation in 2007 (REACH, 2007). REACH requires information on up to 18 physicochemical properties and up to 19 toxicological properties, depending in annual supply levels, of chemicals manufactured in or imported into the European Union. The relevant physicochemical properties for REACH that are amenable to QSPR prediction are: melting/freezing point, boiling point, relative density, vapour pressure, surface tension, water solubility, 1-octanol–water partition coefficient, flash point, flammability, explosive properties, self-ignition temperature, adsorption/desorption, dissociation constant, viscosity, and air–water partition coefficient (Henry's law constant). Dearden et al. (2013) have discussed in detail the QSPR approaches to each of these. An e-book (Benfenati, 2012) gives guidance on the use of QSAR methods for REACH.

The required toxicological properties for REACH include, *inter alia*, skin and eye irritation, skin sensitisation, aquatic toxicity, acute toxicity, sub-chronic rodent toxicity, mutagenicity, and carcinogenicity. Dearden and Rowe (2015) have discussed the role of artificial neural network (ANN) QSARs in the determination of the required toxicological properties.

A defined applicability domain (AD) is important for predictive purposes. One should not use a QSAR/QSPR model to predict an activity/property of a compound that is not reasonably similar to those used to develop the model, otherwise one would have low confidence in the prediction. Netzeva et al. (2005) have examined a number of methods for the determination of AD, and Roy et al. (2015) have recently proposed a new simple approach for that purpose.

In recent years more QSAR work is focussing on mechanistically based modelling, not necessarily because of the OECD guidelines, but because it is important to know the mechanistic basis of a toxic effect, as an aid to developing an Adverse Outcome Pathway. An AOP is a sequential chain of causally linked events at different levels of biological organisation that lead to an adverse health or ecotoxicological effect. AOPs are the central element of a toxicological knowledge framework being built to support chemical risk assessment based on mechanistic reasoning. For example, skin sensitisation is an adverse outcome that has been studied intensively, and a number of mechanisms have been identified (Enoch, Madden & Cronin, 2008).

DESCRIPTORS

As knowledge and expertise in QSAR grew, it became apparent that the simple physicochemical properties used in early post-1962 modelling were not always adequate. In particular, whole-molecule properties became more widely used, allowing modelling of non-congeneric series to be carried out. One of the basic tenets of QSAR modelling is that the compounds used to develop a QSAR should preferably act by the same mechanism of action, otherwise a good model may not be obtained. However, it is not always easy to determine mechanisms, so an assumption was made that compounds within a congeneric series would be more likely to have the same mechanism of action – which is often but not always true. Even within a congeneric series, simple descriptors are not always adequate. For example, the energy of the lowest unoccupied molecular orbital (E_{LUMO}) is frequently used in modelling the aquatic toxicity of electrophiles (see equation 23).

Over the years, many different types of descriptor have become available, and the QSAR practitioner now has many thousands from which to choose. They include simple physicochemical descriptors, those from quantum mechanics (e.g. atomic charge, E_{LUMO}), and those based on molecular topology, such as molecular connectivities (Kier & Hall, 1976, 1986), electrotopological state indices (Hall, Mohney, & Kier, 1991), and graph-theoretic indices (Basak, Niemi & Veith, 1990).

Nowadays descriptors are generally classified as 2D or 3D. 2D descriptors are those that do not depend on molecular conformation; they include most physicochemical descriptors such as log P, molecular weight, and those calculated from graph theory, such as molecular connectivities. 3D descriptors are sensitive to conformation, and include, for example, quantum chemical descriptors, and molecular surface area.

A very interesting set of 5 descriptors are the so-called solvatochromic descriptors, developed by Kamlet et al (1983). They represent polarity, polarizability, hydrogen bond donation, hydrogen bond acceptance, and molecular size, and have been used to model a wide range of both physicochemical and biological properties. It is pertinent to examine here the QSPR for octanol-water partition coefficient (log P) published by Abraham, Chadha, Whiting, and Mitchell (1994):

$$\log P = 0.088 + 0.562\,R - 1.054\,\pi + 0.034\,\alpha - 3.460\,\beta + 3.814\,V_x \qquad (25)$$

$$n = 613\ r^2 = 0.995\ s = 0.116\ F = 23161$$

where R = excess molar refraction (a measure of polarizability), π = polarity/polarizability, α = hydrogen bond donor ability, β = hydrogen bond acceptor ability, and V_x = McGowan characteristic volume. Since the solvatochromic descriptors are approximately auto-scaled, the magnitude or each coefficient indicates the contribution of each descriptor to log P. It can be seen that H-bond acceptor ability and molecular size are the dominant contributors to log P. The former reflects the very strong H-bond donor ability of water, and the latter indicates that more energy is required to create a cavity in water than in octanol, because of strong water-water hydrogen bonding, hence larger molecules favour octanol.

An unusual application of non-molecular descriptors is in ecotoxicology. Although many QSARs that model ecotoxicity use the standard molecular descriptors, Lithner (1989) found that the reproductive toxicity of metal ions to *Daphnia magna* correlated well with the background concentrations of the metal ions in lake water (Lake Superior, USA):

$$\text{Log } (TOX_{rep}) = 0.95 \log (BACKGR) + 1.95 \quad (26)$$

$$n = 22\ r^2 = 0.87$$

where TOX_{rep} = toxic effect level, and BACKGR = typical background concentration of metal ion in oligotrophic lake water.

Walker et al. (2007) developed QSARs for the growth inhibition by 17 metal ions of *Helianthus annuus* 'Sunspot' (sunflower). Using physico-chemical descriptors, they obtained:

$$\log (1/EC_{50}) = 0.636 + 0.000209\ \rho - 0.00367\ \Delta H_s - 0.242 (\log K_1) \quad (27)$$

$$n = 17\ r^2_{adj} = 0.81\ s = 0.265\ F = 24.1$$

where ρ = metal element density, ΔH_s = enthalpy of formation of metal sulfide, and K_1 = stability constant of metal ion with sulfate.

Using background occurrence levels as descriptors, they obtained:

$$\log (1/EC_{50}) = 1.04 - 0.302 (\log M_{soil}) + 0.000014\ X - 0.000025 (\text{Land plants}) \quad (28)$$

$$n = 17\ r^2 = 0.83\ s = 0.25\ F = 27.6$$

where M_{soil} = average metal concentration in soil, X = median elemental composition of soil, and Land plants = mean elemental content in land plants.

Thus the use of background occurrence levels demonstrates their value in ecotoxicity modeling, for they yielded a better correlation than did the use of physico-chemical descriptors. In general, higher background levels mean lower toxicity.

Large compilations of descriptors are now available (e.g. CODESSA, ADAPT, MOE, DRAGON, HYBOT, Molconn-Z).

With so many descriptors now available, how does one choose those that best model the property under investigation? The first step is to eliminate those descriptors with the same value for all compounds, and then to delete one of each pair with high pair-wise collinearity. The rationale for that is that two such descriptors contribute much the same information, and also that collinearity can distort the statistics of a QSAR (Dearden, Cronin, & Kaiser, 2009).

A number of variable selection methods are available (González, Terán, Saíz-Urra, & Teijeira, 2008). The simplest is probably that of step-wise regression, whereby the best single descriptor is selected first, then the next best which, *together with the first*, gives the best two-descriptor model, and so on. The method is fast, but a drawback is that the best n-descriptor model developed in this way is not necessarily the best overall model, since the latter may not incorporate one or more of the descriptors selected by step-wise regression.

Best sub-sets selection eliminates this problem, but is computationally very expensive and time-consuming; it is generally not recommended for pools of more than 20 descriptors.

A very useful variable selection method is the use of genetic algorithms, which imitate genetic selection (Devillers, 1996). It is relatively fast, and has the advantage that it can provide not only the best n-descriptor QSAR, but also the next several best, which allows the selection of a good QSAR

model with appropriate (e.g. easily interpretable) descriptors. A similar method is particle swarm optimization, based on the behavior of bird flocking or fish schooling (Hamzeh-Mivehroud, Sokouti, & Dastmalchi, 2015).

STATISTICS

In order to be of use, a QSAR model must now have appropriate measures of goodness-of-fit, robustness and predictivity (OECD, 2007). Early QSAR work reported few statistics, and occasionally none. Typically QSARs were reported with a correlation coefficient (r) and a standard error of prediction (s) or other measure of error such as root mean square error (RMSE). That gave an indication of goodness of fit, but not of robustness or predictivity. One way to assess robustness is to use an internal validation procedure such as leave-one-out, whereby one compound is removed, and the QSAR is re-developed on the remaining compounds and the correlation coefficient is re-calculated. The removed compound is then returned and a second one removed, and the procedure is repeated, and so on until all compounds have been removed one by one, and an overall correlation coefficient q is calculated. It should be noted that there is still some controversy regarding the value of q, at least so far as its predictive ability is concerned (Golbraikh & Tropsha, 2002; Gramatica, 2007).

Predictivity is vitally important, for one of the key requirements of a QSAR/QSPR model is that it should be able to make reasonably accurate predictions of the requisite activity/property for compounds not used in the development of the model. It is now widely accepted that predictivity of a QSAR should be carried out by means of external validation. That is, the QSAR is used to predict the activities of a test set comprising compounds that are similar to those used in the training set, but were not used in the training set. This is usually performed by splitting a data set into two, a training set (with which the QSAR will be developed), and a test set, typically in an 80%/20% ratio. It is also recommended (Gramatica, 2007) that external validation should not be carried out on QSARs developed using small data sets (< 25 compounds).

Additional useful statistics are: r^2 adjusted for degrees of freedom, which allows comparison of QSARs with different numbers of descriptors; standard errors of the coefficient of each descriptor, which gives an indication of the valid inclusion of a descriptor in a QSAR; and the Fisher statistic (*F*), which gives an indication of the probability that a QSAR is a chance correlation. Dearden et al. (2009) have discussed these statistics in detail.

The value of using standard errors of coefficients is demonstrated by the following QSAR, which models the selectivity of a series of anti-ischaemic tetraalkylbispidines in the rat (Schön, Anton, Brückner, Messinger, Franke, & Gruska, 1998):

$$\log(\text{selectivity}) = 0.37(\pm 0.33)\,MR_1 - 0.010(\pm 0.007)\,(MR_1)^2 + 0.17(\pm 0.10)\,(MR_{3,4})$$

$$- 0.0043(\pm 0.002)\,(MR_{3,4})^2 + 0.43(\pm 0.40)\,I2 - 3.03 \qquad (29)$$

$$n = 16\ r^2 = 0.950\ s = 0.194\ F = 38.3$$

where MR_n = molar refractivity of substituent at position n, and I2 = indicator variable for unsaturation in substituent R_1. Molar refractivity is essentially a volume term, but has a polarizability component. It can be seen that the standard errors of the coefficients of several of the descriptors in equation 29 are almost as large as the value of the coefficient itself, which casts doubt on the validity of those descriptors. In addition, the QSAR does not comply with the Topliss and Costello rule (1972) (*vide infra*).

APPROACHES TO QSAR/QSPR MODELLING

With the availability of so many descriptors, and because *a priori* one often does not know which descriptor(s) will give a good model, many QSAR practitioners start with a large pool of descriptors, and use various techniques, such as eliminating those with high pair-wise collinearities, to reduce the number. Nevertheless it is still necessary to select those few descriptors that will give the best model, and to employ techniques such as step-wise regression, best sub-sets selection or genetic algorithms to do so.

Topliss and Costello (1972) showed that if the ratio of the number of compounds in the training set used to develop a QSAR to the number of descriptors in the QSAR was less than five, there was a significant risk that the correlation was by chance and was not valid. Topliss and Edwards (1979) later showed that, if descriptors were selected from a very large pool, that also increased the risk that chance correlations could arise. It is a matter of concern that both guidelines are still often ignored.

Guidelines for the development of QSAR/QSPR models have been published (Walker, Jaworska, Comber, Schultz, & Dearden, 2003; Walker, Dearden, Schultz, Jaworska, & Comber, 2003; Livingstone, 2004).

In fact, many guidelines for QSAR development and validation are ignored or used incorrectly, which prompted Dearden et al. (2009) to report that they had found 21 different types of error in published QSAR papers, including, *inter alia*, incorrect chemical structures, duplication of compounds, errors in descriptor and end-point values, inadequate data, inadequate or incorrect statistics, unacknowledged omission of data points, over-fitting of data, and incorrect validation.

Other QSAR Approaches

The relatively simple approach in which a set of biological activities is correlated with one or more descriptors is termed multiple linear regression (MLR). It is still widely used, but has a number of drawbacks:

1. It cannot handle non-rectilinear relationships between biological activity;
2. It is very sensitive to outliers;
3. The descriptors should be independent of each other.

Hence, over the years numerous other modelling techniques have been employed in QSAR/QSPR analysis. Many of these were developed originally for purposes other than QSAR analysis.

Artificial neural networks (ANNs), so-called because they are conceptually similar to neural networks in organisms, are fairly extensively used in QSAR/QSPR modelling (Dearden & Rowe, 2015). They have the advantage they can incorporate non-rectilinear functions, such as reciprocal and squared terms. Their disadvantages are that they can easily be over-trained, which would reduce their predictivity, and that they do not yield a QSAR/QSPR equation.

Radial basis function networks are a variant of ANNs. Modarresi, Modarress, and Dearden (2007) used them to develop a QSPR model for Henry's law constant (air-water partition coefficient).

Fuzzy ARTMAP is a synthesis of adaptive resonance theory (ART) neural networks and fuzzy logic. An example of its use in QSAR was for the prediction of aqueous solubility of organic compounds (Yaffe, Cohen, Espinosa, Arenas, & Giralt, 2001). It should be noted that in that work the standard error of prediction was given as 0.08 log unit, whereas the standard error of measurement of aqueous solubility is about 0.6 log unit. Livingstone (1995) has pointed out that in QSAR modeling, if the standard error of prediction is less than that of measurement, over-fitting of the model has taken place.

Principal components analysis (PCA) is a technique that uses an orthogonal transformation to convert a set of values of possibly correlated variables (descriptors) into a set of values of uncorrelated variables called principal components. The transformation is defined so that the first PC has the largest possible variance (i.e. accounts for as much of the variability in the data as possible), and

each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. It is thus useful in effectively reducing a large descriptor pool to manageable proportions.

Discriminant analysis is one technique that can be used for categorical or classification data, for example active/inactive (1/0), or strong/moderate/weak/inactive. In effect it places a multi-dimensional hyperplane between classes, although it is often found that class discrimination is less than 100%. It may be thought of as semi-quantitative, but since the descriptor values used are quantitative, it is classed as a QSAR/QSPR technique.

Logistic regression is similar to discriminant analysis; it estimates probabilities using a logistic function. Worth and Cronin (2003) have discussed the use of both techniques in the development of classification models for human health effects.

Barratt (1995) utilised both PCA and discriminant analysis in a study of the skin corrosivity of organic acids. From the descriptors log P, molecular volume, melting point and pKa, he obtained two PCs that discriminated between corrosive and non-corrosive organic acids (Figure 5).

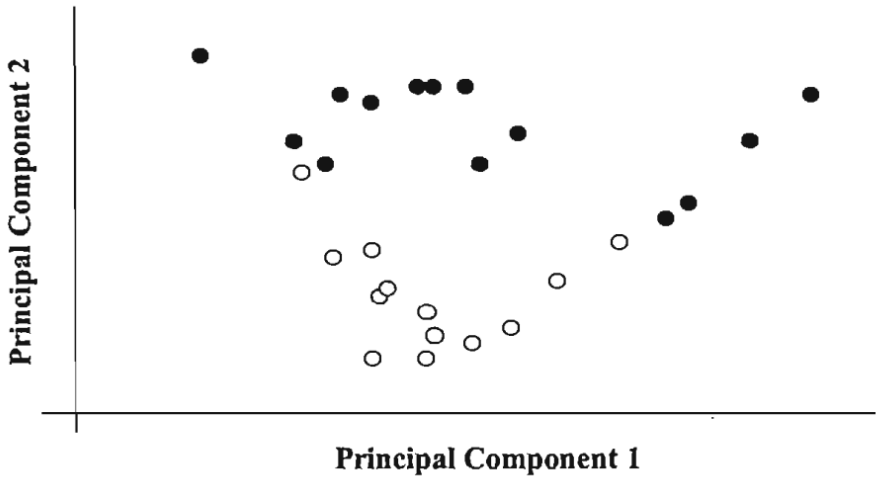
Correspondence analysis is conceptually similar to principal components analysis, but applies to categorical rather than continuous data. Cronin and Dearden (1997) used it to determine the skin sensitisation potential of organic compounds.

Support vector machines (SVMs) are supervised learning models with associated learning algorithms that recognise patterns for classification and regression analysis. They were utilised in the development of QSAR models for predicting anti-HIV-1 activity (Darnag, Mazouz, Schmitzer, Villemin, Jarid, & Cherqaoui, 2010). They allow for more convoluted hyperplane boundaries than does simple discriminant analysis.

Adaptive least squares uses a single discriminant function to make decisions for ordered group discrimination. It has the advantage that it can simultaneously consider any number of classes, and thus can be used for both classification and regression; it is also useful in dealing with data involving activities at only one or two concentrations, when an ED_{50} or LD_{50} cannot be calculated. Schaper and Saxena (1991) have given several examples of its promising performance in regression analysis.

Cluster analysis is used to classify data into groups or categories, based on chemical and/or biological properties, prior to QSAR analysis. It can therefore be used as a data-reduction tool. It is useful when a data set includes chemicals with different mechanisms of action. Yuan and Parrill (2005) used it in the 3D-QSAR analysis of HIV-1 integrase inhibitors.

Figure 5. Discrimination of skin corrosive (o) and non-corrosive (•) organic acids (after Barratt, 1995)



Factor analysis is another data-reduction method, which can also be used for variable selection. Roy and Ghosh (2004) used it for the latter purpose in the modeling of the toxicity of phenylsulfonyl carboxylates to *Vibrio fischeri*.

Partial least squares analysis (PLS) extracts orthogonal (non-correlated) linear combinations of descriptors that explain variances in both descriptors and responses. It is useful when x-variables (descriptors) are correlated, when y-variables (responses) are correlated, and when the number of x-variables greatly exceeds the number of y-variables; this last is frequently the case in modern QSAR analysis, because of the large number of molecular descriptors now available to QSAR practitioners. The method has been well described by, *inter alia*, its developer, H. Wold (Wold, Ruhe, Wold, & Dunn, 1984).

Canonical correlation analysis (CCA) can be used if there are correlations among variables, both dependent (responses) and independent (descriptors). It finds linear combinations of variables with maximum correlation, and can be considered the equivalent of PCA and PLS analysis. Livingstone et al. (2001) used it to predict simultaneously both aqueous solubilities and octanol-water partition coefficients of organic chemicals.

The k-nearest neighbours (k-NN) algorithm is a non-parametric method (i.e. descriptors are not required) used for both classification and regression. The input consists of the (k) closest training set examples. In regression, the output is the property value of the query compound as an average of the values of its k nearest neighbours. Raevsky et al. (2014) used a k-NN approach to model aqueous solubilities of a large number of unionised organic compounds.

Classification and regression trees (CART) is a hierarchical method that determines a consecutive set of *if-then* conditions that allow accurate prediction or classification. Graphically the method resembles a tree, with branching at the *if-then* nodes. It has a number of attractive features, including simplicity, interpretability, capacity to handle large data sets and model non-linearities, no assumptions regarding data distribution, immunity to outliers and collinearities. Tan et al. (2010) have used the method to model the antimicrobial potencies of diverse agents against *Candida albicans*.

Random forests are an extension of CART; they are ensemble learning methods that use a number of decision trees, the output being the mode of the classes in classification modeling, and the mean prediction of the individual trees in regression modeling. Random forests correct for the overfitting that can occur with decision trees. An example of their use in QSAR is the prediction of toxicity of diverse organic chemicals to the aquatic ciliate *Tetrahymena pyriformis* (Polishchuk, Muratov, Artemenko, Kolumbin, Muratove, & Kuz'min, 2009).

SIMCA (Soft Independent Modeling of Class Analogy) was developed to enhance the classification potential of principal component models. It constructs separate models for different classes, and new objects (chemicals) of unknown class can be projected into the same multivariate space as the training set. It has the advantage of being able to handle large numbers of collinear variables. Giaginis et al. (2014) used it for multivariate data analysis in the QSAR modeling of redistribution of antidepressants.

Mapping techniques such as Kohonen mapping and non-linear mapping are visualization methods; they map a set of vectorial samples onto a two-dimensional lattice whilst preserving the topology of the original space. The resulting patterns can serve as indicators of structure-activity relationships, and can also aid in classification. Polanski et al. (2000) used Kohonen mapping to aid in the selection of new artificial sweetener candidates. Non-linear mapping in QSAR and QSPR has been reviewed by Domine et al. (1993).

Consensus modelling is an approach used in both drug research and toxicology; it involves taking an account, for example by averaging, of predictions from a number of QSAR and/or other models. Matthews et al. (2008) combined predictions from five software programs to improve the prediction of chemical carcinogenesis in rodents. Mitra et al. (2012) used consensus modelling to predict drug-induced adverse reactions.

An important 3D-QSAR approach is comparative molecular field analysis (CoMFA), developed by Cramer et al. (1988). It allows comparison of molecules by aligning them in space and mapping their steric and electrostatic fields on a 3-D grid. It overcomes the usual QSAR problem of an excessive number of descriptors (molecular field values) by the use of partial least squares statistics. CoMFA requires molecular alignment, which can be difficult, and the electrostatic potentials used are very steep at the van der Waals surface, meaning that the potential energy changes rapidly. A variation of CoMFA is CoMSIA (comparative molecular similarity index analysis), developed by Klebe and Abraham (1999), which includes hydrogen bonding as well as steric and electrostatic fields. The contour maps obtained are better than those from CoMFA, in that they are easy to interpret and very intuitive as a visualization tool. da Cunha et al. (2009) used both CoMFA and CoMSIA to generate 3D-QSAR predictions of the activity of some HIV-1 protease inhibitors. Yuan and Parrill (2005) used the related technique of molecular field analysis (MFA) to model the activity of a series of HIV-1 integrase inhibitors.

A similar approach is GRID, which uses a smoother potential function than the Lennard-Jones type used by CoMFA. It explores not only steric and electrostatic potentials, but also hydrogen bonding potential, using a variety of probes. Kim et al. (1993) found it to be superior to CoMFA in a study of benzodiazepines.

Molecular shape analysis involves the characterization and use of molecular shape in the development of a QSAR model (Roy, Kar, & Das, 2015b). It has the advantage that knowledge of receptor geometry is not required. Rowberg et al. (1994) used it in the QSAR modelling of 1-(phenylcarbamoyl)-2-pyrazolin insecticides. Comparative receptor surface analysis (CoRSA) is another technique that does not require knowledge of receptor geometry. It generates a virtual receptor represented as points on a surface. The CoRSA descriptors are then used in, for example, a PLS analysis to construct a QSAR model. It was used by Ivanciuc et al. (2001) to model the calcium channel antagonist activity of some dihydropyridine derivatives.

Common reactivity pattern analysis (COREPA) uses a Bayesian probabilistic method to identify common structural characteristics among chemicals that elicit similar biological activity or class in a context that allows many possible conformations of individual chemicals and the probability distribution of molecular descriptor values instead of single parameter values for each chemical. The common structural characteristics can then be encoded into a decision tree to screen large and structurally heterogeneous chemical libraries for the required biological activity (Mekenyan et al., 1999).

In the fragment-based QSAR (FB-QSAR) approach the bioactivities of molecules are correlated with the physicochemical properties of molecular fragments through two sets of coefficients in linear free energy equations. One coefficient set is for the physico-chemical properties and the other for the weight factors of the molecular fragments. Then an iterative double least square technique is developed to solve the two sets of coefficients in a training data set alternately and iteratively. The FB-QSAR approach can remarkably enhance the predictive power and provide more structural insights into rational drug design. Du et al. (2009) used it to build a predictive model of neuraminidase inhibitors for drug development against H5N1 influenza virus. Group-based QSAR (G-QSAR) is an extrapolation of FB-QSAR that allows substituent interactions as fragment-specific descriptors to account for fragment interactions (Ajmani, Jadhav, & Kulkarni, 2009). It has the advantage that it highlights the molecular site(s) where optimization is required in drug design. This means that it can deal with the inverse QSAR problem; that is, it offers a method for the design of new molecules from the QSAR model. There have, of course, been a number of other studies of inverse QSAR (e.g. Wong & Burkowski, 2009; Hasegawa, Kimura & Funatsu, 2009).

Another important fragment-based QSAR technique that does not require molecular alignment is hologram-based QSAR (HQSAR). In this approach fragments are assigned integer values, and these are then used to make a fixed-length integer array, termed a molecular hologram; the bin occupancies of the hologram are used as descriptors. The approach encodes all possible fragments. PLS is then

used to build a QSAR model. In a comparative study of CoMFA, CoMSIA and HQSAR in modeling acetylcholinesterase inhibitors, Jiang et al. (2013) found HQSAR superior to the other techniques based on internal (q^2) validation, but intermediate based on external validation.

Despite the vast number of published papers, edited books and book chapters on QSAR, there are but a few books that can be called QSAR text-books, with basic information and practical advice on how to develop, validate and use QSARs. The first such book was by Yvonne Martin (1978), followed over 30 years later by a second edition (Martin, 2010). Others were by Raevsky (1984) (published in Russian), and Kubinyi (1993); the latest are those of Roy et al. (2015a, 2015b).

REFERENCES

1. Todeschini, Roberto; Consonni, Viviana. Handbook of Molecular Descriptors. Wiley. ISBN 978-3-527-29913-3.

2. Mauri, Andrea; Consonni, Viviana; Todeschini, Roberto. "Molecular Descriptors". Handbook of Computational Chemistry. Springer International Publishing. pp. 2065–2093. ISBN 978-3-319-27282-5.

3. Roberto Todeschini and Viviana Consonni, Molecular Descriptors for Chemoinformatics (2 volumes), Wiley-VCH, 2009.

4. Mati Karelson, Molecular Descriptors in QSAR/QSPR, John Wiley & Sons, 2000.

5. James Devillers and Alexandru T. Balaban (Eds.), Topological indices and related descriptors in QSAR and QSPR. Taylor & Francis, 2000.

6. Lemont Kier and Lowell Hall, Molecular structure description. Academic Press, 1999.

7. Alexandru T. Balaban (Ed.), From chemical topology to three-dimensional geometry. Plenum Press, 1997