SAR and QSAR in Environmental Research

ISSN: 1062-936X (Print) 1029-046X (Online) Journal homepage: http://www.tandfonline.com/loi/gsar20

How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR)

J.C. Dearden , M.T.D. Cronin & K.L.E. Kaiser

SAR and QSAR in Environmental Research
Vol. 20, Nos. 3–4, April–June 2009, 241–266

Taylor & Francis
Taylor & Francis Group

How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR)

J.C. Dearden[a]*, M.T.D. Cronin[a] and K.L.E. Kaiser[b]

[a]*School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK;* [b]*TerraBase Inc., 1063 King Street West, Suite 130, Hamilton, Ontario L8S 4S3, Canada*

(*Received 28 November 2008; in final form 30 March 2009*)

Although thousands of quantitative structure–activity and structure–property relationships (QSARs/QSPRs) have been published, as well as numerous papers on the correct procedures for QSAR/QSPR analysis, many analyses are still carried out incorrectly, or in a less than satisfactory manner. We have identified 21 types of error that continue to be perpetrated in the QSAR/QSPR literature, and each of these is discussed, with examples (including some of our own). Where appropriate, we make recommendations for avoiding errors and for improving and enhancing QSAR/QSPR analyses.

Keywords: QSAR/QSPR; OECD principles; errors; data; descriptors; statistics

1. Introduction

Quantitative structure–activity and structure–property relationships (QSARs/QSPRs), as currently understood, have been developed and used for close to 50 years, starting with the seminal work of Corwin Hansch and co-workers on pesticides in 1962 [1]. If we consider QSPRs, the first was probably that of Mills [2] on prediction of melting and boiling points in homologous series. His QSPR for melting points was

$$\text{M.P.} = \beta(x - c)/(1 + \gamma(x - c)) \tag{1}$$

where x is the number of CH_2 groups in the chain and β, γ and c are constants. Mills found that for many series of compounds, melting points could be predicted to within 1–2°.

Since 1962 many thousands of QSARs and QSPRs have been developed, covering a vast range of endpoints, and various statistical techniques have been used, both to develop the correlations and to assess their goodness of fit and validity. To that end, a number of publications have offered guidance on the correct procedures in QSAR/QSPR development [3–9], in addition to numerous books on the subject.

The primary value of a QSAR or QSPR is its predictivity, that is, how well it is able to predict endpoint values of compounds not used to develop the correlation, i.e. not in the training set. Two main methods are used to determine predictivity: internal cross-validation and external validation with a test set of compounds. It is generally accepted now that only QSARs and QSPRs that have been suitably externally validated can be considered dependable for both scientific and regulatory purposes [10,11].

*Corresponding author. Email: j.c.dearden@ljmu.ac.uk

ISSN 1062–936X print/ISSN 1029–046X online

DOI: 10.1080/10629360902949567
http://www.informaworld.com

In March 2002 a meeting of QSAR/QSPR experts was held in Setúbal, Portugal, to formulate a set of guidelines for the validation of QSARs/QSPRs, in particular for regulatory purposes [12]. Six guidelines were drawn up, which were later adopted by the Organisation for Economic Co-operation and Development (OECD) [13] and modified to five. The guidelines are that a valid QSAR/QSPR should have:

(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.

The guidelines are now known as the OECD Principles for the Validation of (Q)SARs, although they are intended to apply also to QSPRs. The OECD has also provided a checklist to provide guidance on the interpretation of the principles [14]. It is interesting to note that similar guidelines were proposed as long ago as 1973 [15].

Despite the foregoing, errors continue to be made in the development and use of QSARs and QSPRs, and this paper examines the main types of error, with examples.

2. Errors in QSARs/QSPRs

Over 20 main types of error can be found, and continue to be made, in QSAR/QSPR development and use, and these are listed in Table 1, together with the corresponding OECD principle(s).

Table 1. Types of error in QSAR/QSPR development and use.

No.	*Type of error*	*Relevant OECD principle(s)*
1	Failure to take account of data heterogeneity	1
2	Use of inappropriate endpoint data	1
3	Use of collinear descriptors	2, 4, 5
4	Use of incomprehensible descriptors	2, 5
5	Error in descriptor values	2
6	Poor transferability of QSAR/QSPR	2
7	Inadequate/undefined applicability domain	3
8	Unacknowledged omission of data points	3
9	Use of inadequate data	3
10	Replication of compounds in dataset	3
11	Too narrow a range of endpoint values	3
12	Over-fitting of data	4
13	Use of excessive numbers of descriptors in a QSAR/QSPR	4
14	Lack of/inadequate statistics	4
15	Incorrect calculation	4
16	Lack of descriptor auto-scaling	4
17	Misuse/misinterpretation of statistics	4
18	No consideration of distribution of residuals	4
19	Inadequate training/test set selection	4
20	Failure to validate a QSAR/QSPR correctly	4
21	Lack of mechanistic interpretation	5

2.1 *Failure to take account of data heterogeneity*

It is important that, within a given dataset, all endpoint values are consistent [16]. For example, there are several ways of determining aqueous solubility: in pure water, as undissociated species (intrinsic solubility), at a given pH, and at a given ionic strength. There are also a number of different methods for the determination of solubility, each of which could conceivably yield slightly different results. In addition, temperature clearly affects solubility. For best results, then, a set of solubility values should have been determined using the same protocol. In general, for data collected from the literature, this is unlikely, so that a QSPR developed from such data will not be as accurate as one developed from single-protocol data.

Another aspect of this problem relates to biological data determined by different protocols. Again, this tends to occur when data are collected from different literature sources. For example, the modelling of hERG potassium ion channel inhibition used data obtained with human embryonic kidney cells and with Chinese hamster ovary cells [17]. A QSAR with a squared correlation coefficient (r^2) of 0.70 and a standard error (s) of 0.76 was obtained with 104 compounds. By contrast, Coi et al. [18], using only data from human embryonic kidney cells, obtained a QSAR with $r^2 = 0.77$ (s not given) with 55 compounds. The improvement is probably due at least in part to the use of a single cell line by Coi et al. [18].

Data for compounds from different chemical classes are often incorporated into QSAR/QSPR correlations. Strictly, this is acceptable, provided that the mechanism of action of all of the compounds is the same, and there are many published QSARs and QSPRs developed with training sets containing diverse chemical classes. It is fair to say that the accuracy of prediction is usually less for such correlations, partly because the data are drawn mostly from published literature, and hence have been determined in different laboratories. Of course, data for compounds within a single chemical class can also have been determined in more than one laboratory. It should be noted that there are different mechanisms of action involved in physico-chemical processes. For example, in octanol–water partitioning the mode of action is the same for all chemicals, but the mechanisms can differ. For example, a hydrocarbon cannot partition via hydrogen-bonding ability, whilst for, say, aliphatic alcohols, hydrogen bonding is an important partitioning mechanism.

One of the present authors had occasion a number of years ago to review a manuscript that reported the development of a QSAR based on eight related compounds, whose activities had all been measured in the same laboratory. The authors found that the activities could be modelled well by a quadratic equation in $\log P$, with $r^2 = 0.906$ (Figure 1). However, examination of the data revealed that whilst the six compounds on the left of Figure 1 (●) were aliphatic, the two on the right (○) were aromatic. The inclusion of all eight compounds in the same QSAR was therefore probably not justified, and the manuscript was rejected. More aromatic derivatives would be needed to establish the true relationship(s).

2.2 *Use of inappropriate endpoint data*

One of the commonest, and most severe, errors in QSAR/QSPR modelling is to use data with incorrect units, especially the use of concentrations or doses in weight rather than in molar units. The effect (biological or physico-chemical) of a chemical is a consequence of

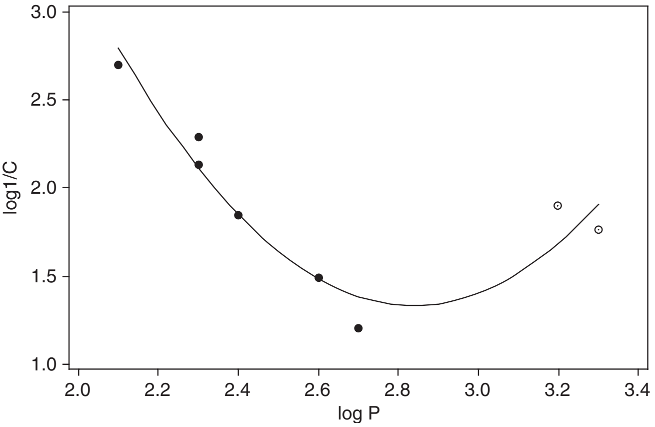

Figure 1. Incorrect fitting of a quadratic equation to heterogeneous data.

the number of molecules present, and not how much they weigh (although molecular size can, of course, be a factor in, say, membrane penetration). It is particularly difficult to get this point across to pharmacologists and toxicologists, who are generally most reluctant to change from mg/kg (or mg/L for aquatic systems) to mmol/kg (or mM/L). If the endpoint being measured is one determined from a dose–response relationship (e.g. LD_{50}), it is a simple matter to convert from weight to molar units. If, however, the response is measured at a fixed weight dosage (e.g. 10 mg/kg) for all chemicals, then conversion is useless, because each chemical will have been tested at a different molar dose, and so strict comparison is impossible.

An example of the first type of error is a study of the mammalian toxicity of 115 substituted anilines [19], in which the data used were LD_{50} values in mg/kg. The very weak correlations obtained were undoubtedly due in part to the fact that the toxicity data were taken from the RTECS (Registry of Toxic Effects of Chemical Substances) database [20], which includes many data of variable quality, but the use of LD_{50} values in mg/kg would have contributed to the weakness of the correlations. The same error was made in a study of health effects of chemicals in humans [21]; correction to mmol/kg units improved the correlations [22].

An example of the second type of error is a QSAR study of *in vitro* local anaesthetic activity, in which the activity was measured at 1% w/v for 19 derivatives of 3-aminobenzo-[d]-isothiazole (see [23]). Although in this case there was not a great range of molecular weights among the compounds, the activity was nevertheless measured at a different concentration for each compound, which must have detracted from the accuracy of the QSAR correlation.

It should be noted that *SAR and QSAR in Environmental Research* does not accept papers that report QSARs and QSPRs based on weight concentrations/dosages. It would be good to see other journals following this practice.

2.3 *Use of collinear descriptors*

The use of collinear descriptors is undesirable, for two reasons. First, two highly collinear descriptors are effectively contributing the same information twice, and thus add nothing to mechanistic interpretation of a QSAR or QSPR; in fact, they can detract from it. Second, it is known [5,7] that the use of collinear descriptors can have adverse effects on the statistical analysis, for example by causing instability in the regression coefficients.

One of the easiest ways to fall into the collinearity trap is to use several molecular connectivity terms, the lower orders of which are highly collinear [24]. An example is the modelling of the fish bioconcentration factor (BCF) of nonpolar organic pollutants [25]:

$$\log \mathrm{BCF} = 0.770 + 0.757\,{}^{0}\chi^{\mathrm{v}} - 2.650\,{}^{1}\chi + 3.372\,{}^{2}\chi - 1.186\,{}^{2}\chi^{\mathrm{v}} - 1.807\,{}^{3}\chi_{\mathrm{c}} \tag{2}$$

$$n = 80, \quad r^2 = 0.9066, \quad s = 0.3636$$

Bearing in mind the high measurement error on BCF values [26], it is unlikely that Equation (2) is really able to describe 90% of the variation in log BCF.

Another aspect of collinearity is the use of, say, a single descriptor to model data, and to attribute some mechanistic significance to that descriptor, whilst failing to recognise that a second descriptor, with different mechanistic significance but highly correlated with the first, also models the data well.

In a paper examining the skin sensitisation potential of a series of 1-bromoalkanes [27], the following QSAR was obtained:

$$\log 1/(T/C_3) = 1.61 \log P - 0.09\,(\log P)^2 - 7.4 \tag{3}$$

$$n = 9, \quad r^2 = 0.94, \quad s = 0.11, \quad F = 50$$

However, Cronin and Schultz [6] have pointed out that since, in this series of compounds, $\log P$ and molecular size are totally collinear, the same result could have been obtained using molecular weight or number of carbon atoms as the descriptor. Hence, it cannot be determined whether hydrophobicity or molecular size is the controlling factor. The synthesis of the compounds was clearly not guided by mechanism of action requirements.

A similar example is a study of the toxicity of a series of alkyl ethers to the mouse [28], which yielded a good correlation with first-order molecular connectivity (${}^{1}\chi$), an indicator of molecular size [24]. However, Cronin and Schultz [6] have pointed out that a very similar correlation obtains with $\log P$, because ${}^{1}\chi$ and $\log P$ are collinear in this series ($r^2 = 0.98$).

An easy way to check whether collinearity is affecting a correlation is to look at the probability (p) values of each coefficient. Using the Randić and Basak data [28], the QSAR with ${}^{1}\chi$ as descriptor is

$$\log 1/C = 0.634\ (\pm 0.066)\,{}^{1}\chi + 0.789\ (\pm 0.197) \tag{4}$$

$$n = 21, \quad r^2 = 0.828, \quad s = 0.176, \quad F = 91.2, \quad p({}^{1}\chi) < 0.001$$

Using Randić and Basak's weighted identification (WID) number as descriptor, the QSAR is

$$\log 1/C = 0.325\ (\pm 0.027)\ \mathrm{WID} + 0.602\ (\pm 0.169) \tag{5}$$

$$n = 21, \quad r^2 = 0.887, \quad s = 0.143, \quad F = 148.8, \quad p(\mathrm{WID}) < 0.001$$

Now ${}^{1}\chi$ and WID are highly collinear ($r^2 = 0.959$), and if both descriptors are incorporated into the correlation, the resulting QSAR is

$$\log 1/C = 0.426\ (\pm 0.133)\ \mathrm{WID} - 0.208\ (\pm 0.268)\ {}^{1}\chi + 0.577\ (\pm 0.174) \qquad (6)$$

$$n = 21, \quad r^2 = 0.890, \quad s = 0.144, \quad F = 73.2, \quad p(\mathrm{WID}) = 0.005, \quad p({}^{1}\chi) = 0.447$$

Despite the fact that both WID and ${}^{1}\chi$ give good correlations with log $1/C$, in equation (6) ${}^{1}\chi$ appears not to be a good descriptor, because both the standard error of its coefficient and its *p*-value are very high. Furthermore, interpretation of the QSAR is distorted because ${}^{1}\chi$ has a negative coefficient in Equation (6), despite having a positive coefficient in Equation (4), which is not uncommon if collinearity is present [29].

2.4 *Use of incomprehensible descriptors*

Thousands of molecular descriptors are now available [30], and many of them are very difficult, if not impossible, of clear physico-chemical interpretation [31]. It behoves us, using the principle of Occam's razor (see Section 2.13), to keep QSARs and QSPRs simple both qualitatively and quantitatively [15].

There are three ways in which incomprehensibility can happen. First, mention of the names and numbers of descriptors can be omitted altogether. This can happen when a modelling technique such as an artificial neural network is used that does not yield an equation. An example is a study of oestrogen binding using CODESSA descriptors [32] in which the descriptors selected for the QSAR are not reported. Second, some or all of the descriptors can have such abstruse meanings that their physicochemical relevance is unclear or unknown [19,33,34]. Third, explanations of the meanings of descriptors can be lacking. Owing to the huge number of descriptors, some of them rather esoteric, that can be used in QSAR/QSPR analysis, even experienced practitioners cannot know the meaning of all of them. It is nevertheless important that a reasonable explanation be given of the meaning of the descriptors used in a published QSAR/QSPR. Most authors do give such an explanation, albeit often a very brief one. However, some publications do not do so, to the detriment of readers' understanding of the work. An example is an interesting paper on the characterisation of chemical and biological properties using a range of QSAR/QSPR methods [35]. Its value would have been enhanced considerably by the inclusion of an explanation of the descriptors used.

2.5 *Error in descriptor values*

It is widely recognised that measured property values, be they biological or physico-chemical, will contain error. What is not understood so well is that most descriptor values, measured or calculated, can also contain errors, although a few, such as atom counts and molecular connectivities, are error-free provided that no arithmetic errors have been made in their calculation.

It therefore behoves the QSAR/QSPR practitioner to ensure that the descriptor values used are as accurate as possible. This is not a simple task, as how does one know whether a descriptor value is accurate or not? In the case of properties for which several methods of measurement or calculation are possible, one way is to take an average of several values. For example, Dearden [36] carried out a comparison of 14 commercially available software programs for the prediction of aqueous solubility, using a test set of 122 drugs with accurately measured solubilities [37]. The predictions of the four best-performing programs for three of the drugs are shown in Table 2.

Table 2. Aqueous solubility (log S, with S in mol/L) predictions for three drugs from four well-performing software programs.

Software	*Atropine*	*Caffeine*	*Butylparaben*
WSKOWWIN (http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm)	−1.87	−1.87	−3.09
CSLogWS (http://www.chemsilico.com)	−2.06	−0.65	−3.05
SPARC (http://ibmlc2.chem.uga.edu/sparc)	−1.01	−0.27	−3.07
ADME Boxes (http://www.acdlabs.com)	−2.03	−0.56	−2.58
Mean	−1.74	−0.84	−2.95
Measured	−2.18	−1.02	−2.96

It can be seen that although at least one program yields a poor prediction for each of the three drugs, the mean of the four predictions is well within the mean experimental error of about 0.6 log unit for aqueous solubility [38]. Other consensus approaches have shown similar improvements in prediction [39–41].

The values of most calculated properties depend on the method of calculation, and so there is a need for comparative checks of performance. Some work in this area has been reported [36,42–44], but much remains to be done.

2.6 *Poor transferability of QSARs/QSPRs*

Since one of the prime uses of a QSAR or QSPR is that others can use it for predictive purposes, it is important that it is transferable to other workers and is reproducible by them. Hartung et al. [45] have suggested the following criteria for transferability of a QSAR/QSPR to a different operator:

(a) descriptor values can be reproduced;
(b) model definition can be confirmed;
(c) goodness of fit and statistical robustness can be confirmed;
(d) reproducibility of predictions can be confirmed;
(e) an assessment is given of the adequacy of documentation on the development and application of the model.

Transferability is a relatively new explicit concept (although it is implicit in the publication of every QSAR/QSPR), and there are very few, if any, published QSPRs to date whose authors have reported transferability. Roncaglioni [46] has discussed the criteria for selecting promising QSARs/QSPRs for independent evaluation, which is especially important in a regulatory context. The criteria that she used were valid experimental data, chemical information on structures, chemical information on descriptors, chemical domain, and the modelling approach and feasibility of the model.

That such criteria are frequently not satisfied has been demonstrated a number of times in the present paper. We recommend that the Roncaglioni criteria are borne in mind in every QSAR/QSPR investigation and publication.

2.7 *Inadequate/undefined applicability domain*

The applicability domain (AD) of a QSAR/QSPR has been defined [47] as: ‘the response and chemical structure space in which the model makes predictions with a given reliability'. Netzeva et al. [47] and Jaworska et al. [48] have reviewed methods for defining an AD, and made a number of recommendations for their determination and use. They recommended that the definition of an AD should be the responsibility of the model builder rather than the model user, and that the starting point must be the publication of all of the training set compounds, including both structures and descriptors. Very few QSAR/QSPR papers even now comply with that recommendation. Furthermore, there are currently only a few software programs for property prediction that give an indication of whether or not a test compounds falls within the AD of the training set compounds used to develop the software. However, caution is advised concerning the reliability of predictions even for compounds that lie within the AD boundary. One reason for this could be the existence of so-called 'activity cliffs' [49,50].

Currently very few published QSAR/QSPR papers give any information about the AD, and most use test set compounds that are (qualitatively) reasonably similar to those of the training set. This clearly needs to change, and we recommend that every published QSAR/QSPR should be accompanied by an indication of its AD. If this is not done, then one or more test set compounds could lie outside the AD [51].

We also suggest that the term 'applicability domain' is changed. Its use implies that predictions for compounds outside the AD are invalid, whereas in fact they are simply less reliable. We prefer the term 'optimum prediction space' used in the TOPKAT software [52].

2.8 *Unacknowledged omission of data points*

In compiling a set of chemicals for use in a QSAR/QSPR analysis, data are often taken from published literature, and in many cases only a selected number of data points are taken. There is usually a valid reason for such selection, for example to keep the number of data to a manageable level, or to study a specific chemical class. However, instances occur when data are pruned without any reason being given, or even any acknowledgment made that pruning has been carried out.

In 1995 Dearden et al. [53] published a QSAR study of the toxicity of 47 nitrobenzenes to the aquatic ciliate *Tetrahymena pyriformis*. A recent publication [54] used the same dataset, but omitted, without explanation, eight of the original 47 compounds. There could well have been a valid reason for this omission, for example that the eight compounds were outliers in the QSAR analysis, but it should have been given.

Concerning outliers, Lipnick [55] commented that 'an outlier among residuals is one that . . . perhaps lies three or four standard deviations or further from the mean of the residuals. The outlier is a peculiarity and indicates a data point which is not at all typical of the rest of the data. It follows that an outlier should be submitted to particularly careful examination to see if the reason for its peculiarity can be determined'. Cronin and Schultz [6] have discussed in detail for methods available for the detection and removal of outliers, stating that 'ideally, there should be a valid reason for outlier removal. This (is) normally due to compounds operating by a different mechanism of action, whether this be toxicological or pharmacological, compounds being atypical of the dataset, or other physical effects including steric hindrance of reactive centres (or functional groups required for receptor binding) and ionisation. When performed correctly, removal of significant outliers will allow for the development of stronger and more significant models'. On the other hand, Aptula et al. [56] found that root mean square error of prediction could be lower if all available chemicals in the training set were used for model development, regardless of whether they were statistical outliers.

It is therefore strongly recommended that when pruned data are used, a valid explanation is given for the pruning.

2.9 *Use of inadequate data*

Inadequacy of data can mean several things; some have already been mentioned, such as those in sections 2.1, 2.2, 2.7, and 2.8. Another important potential problem with data is their accuracy, which is often difficult to assess, especially if only one endpoint value is available for a given chemical, or if widely differing endpoint values are given. For example Nendza [26] reported that measured BCFs for pentachlorobenzene ranged from 900 to 250,000.

Another common problem is the inclusion of incorrect or inadequately defined names, chemical structures or CAS numbers in data for QSAR/QSPR analysis. A QSAR study [57] of the central nervous system (CNS) permeability of drugs reported different descriptor values (including molecular weights differing by 14) for aspirin and acetylsalicylate. Even misspelled names of chemicals can cause confusion, for example with cimetidine and ketoprofen [58].

A very recent paper [59] reported that incorrect chemical structures in six public and private databases ranged from 0.1% to 3.4%. It was found that slight errors in chemical structures (e.g. incorrect location of a chlorine atom) can cause significant differences in QSAR/QSPR prediction accuracy.

Isomeric chemicals are frequently not sufficiently defined. For example, a QSAR study of skin absorption [60] listed 4-chlorocresol and chloroxylenol among the dataset used. There are two isomeric 4-chlorocresols and 18 isomeric chloroxylenols.

The use of property values, especially biological activities, in a QSAR/QSPR implies that the values are the most accurate available. Whilst it is acceptable (and, indeed, sometimes necessary) to use calculated (predicted) descriptor values (e.g. log P values) in QSAR/QSPR analysis, it is not acceptable to use predicted values of the property (activity) to be modelled, as then one is making predictions of predicted values. A paper devoted to the prediction of skin permeability [61] used a dataset of 114 compounds, 63 of which had calculated permeability values. The same dataset was also used by ourselves [62], for which we were rightly criticised [16,63].

Some databases, such as RTECS [20] contain unchecked data, which cannot be relied upon for QSAR analysis.

2.10 *Replication of compounds in a dataset*

One of the most common errors found in many QSAR/QSPR publications, even in peer-reviewed journals, is the presence of replicate structures. When identical structures are present in the training set, these replicates may unduly influence the resulting algorithm. When identical structures are present in both the training and testing sets, the resulting statistics can create the impression of a much better predictive power of the presented model than is actually the case. Therefore, the potential presence of replicates in any dataset should be thoroughly checked prior to undertaking any QSAR/QSPR analysis.

Replicate structures can arise for a number of reasons. Primarily, they include different chemical names [64], different CAS registry numbers, different structure codes (e.g. SMILES strings), different activity or property values for the same compound [57], different numbering systems of chemicals, and even simple repetition, as is the case with a recent paper concerning simulation of nuclear magnetic resonance (NMR) spectra [65].

Most compounds have more than one name, and some have many (excluding trade names). Table 3 shows the names listed by ChemFinder [66] for two widely used compounds.

Perhaps surprisingly, many compounds have more than one CAS registry number. For example, cyanoguanidine has no fewer than nine [67]: 461-58-5, 125148-58-5, 139351-77-2, 139351-78-3, 157480-33-6, 166432-96-8, 187414-06-8, 205265-14-1, and 313058-80-9.

Structure codes can also be presented in different ways. SMILES strings, for example, can generally be written in a multiplicity of ways, although it is possible to write unique SMILES strings [68].

Table 3. Some synonyms of tetracycline and acrolein (from [64]).

Tetracycline
Aureocarmyl;
Aureomycin;
Aureomycin A-377;
Aureomycin-R;
Aureomykoin;
Aurofac;
Aurofac 10;
Biomitsin;
Biomycin;
Biomycin A;
Chlorotetracycline;
7-Chlorotetracycline;
7-Chloro-4-(dimethylamino)-1,4,4a,5,5a,6,11,12a-octahydro-3,6,10,12,12a-pentahydroxy-6-methyl-1,11-dioxo-2-naphthacenecarboxamide;
Chrysomykine;
CTC;
Duomycin;
Flamycin;
Naphthacenecarboxamide, 7-chloro-4-(dimethylamino)-1,4,4a,5,5a,6,11,12a-octahydro-3,6,10,12,12a-pentahydroxy-6-methyl-1,11-dioxo-, (4S-(4α,4aα,5aα,6β,12aα))-.
Acrolein
Acraldehyde;
Acrylaldehyde;
Acrylic aldehyde;
Allyl aldehyde;
Crolean;
Ethylene aldehyde;
Prop-2-en-1-al;
Prop-2-enal;
Propenal;
Propenaldehyde;
2-Propen-1-one;
Propylene aldehyde;
trans-Acrolein.

Different notation for chemicals can also cause confusion. The current InChI (International Chemical Identifier) atom numbering system developed by IUPAC (International Union of Pure and Applied Chemistry) [69] is widely accepted, but other notation is still in use, such as an earlier IUPAC system used in the older editions of the *CRC Handbook of Chemistry and Physics* [70]. Figure 2 shows the two numbering systems for benz[*a*]anthracene.

Clearly the avoidance of replicates should be of concern when developing and testing QSARs/QSPRs. Most frequently, the presence of replicates results from inadequate or incorrect information on the compounds studied, resulting in unintentional replicates (occasionally with different biological data), identical compounds in training and/or test sets, and resulting faulty statistics. Given the possibilities of replicates arising for a variety of reasons, as indicated above, the question then becomes how to recognise and eliminate such occurrences. The answer to this question lies in the preparation of unique structural codes, such as the InChI code [69], and checking with a computerised system for replicates. For example, modern spreadsheets have built-in replicate recognition and/or removal functions. Failing the availability of unique structural codes, the next best approach is to use a spreadsheet and to sort all data by each of the available parameters, especially chemical formula and biological effect data, for possible replicates. Any compounds with identical sum-formula, but different values, should also thoroughly be checked to ensure that no identical structures are present with different effect values.

With respect to InChI codes, the reader should note that the conversion of SMILES strings to InChI codes is currently available only via an intermediate conversion to the corresponding SDF (Standard Data Format). In our experience, this double conversion produces consistently incorrect InChI codes when dealing with many types of organometallic compounds.

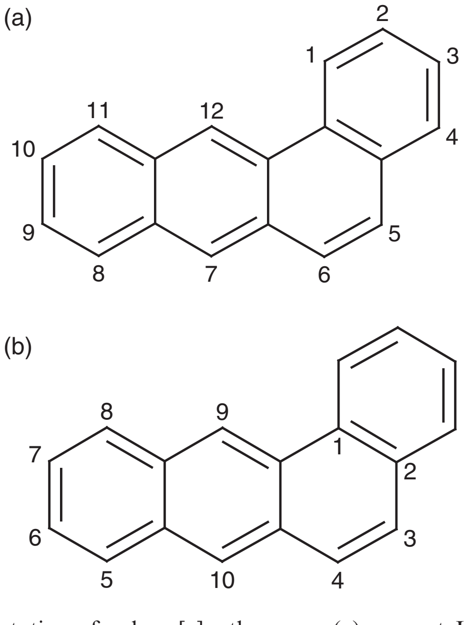

Figure 2. Two different notations for benz[*a*]anthracene: (a) current InChI notation; (b) former IUPAC notation.

The extra work involved in the preparation of a clean, replicate-free dataset at the beginning of QSAR/QSPR work is well worth the effort. It avoids much additional work when the problem is recognised at a later stage of the research and publication process.

In a nutshell, sorting and re-sorting the data by all available means, and checking and re-checking for replicates and other inconsistencies are of paramount importance. Authors should also sort tables in such a way that the understanding of the data and their variation is most readily apparent to the reader. Tables can be ordered alphabetically, by CAS number or by property value, for example. However, quite often the ordering of compounds in a table appears to be random, as is the case with a QSPR study of soil sorption [71], in which compound names, CAS numbers and log K_{oc} values are tabulated haphazardly.

2.11 *Too narrow a range of endpoint values*

If the endpoint values of a set of compounds are all identical, a QSAR/QSPR model cannot be developed. It thus appears that with a very narrow range of endpoint values, it would be difficult to develop a good model, and so the greater the range of endpoint values, the better can compounds be modelled. This is in fact the case, as has been demonstrated by Gedeck et al. [72], who showed, using a number of different datasets, that predictive r^2 values improve steadily with the increase in the range of endpoint values. They concluded that a range of endpoint values of at least 1.0 log unit (a 10-fold range of actual values, e.g. LD_{50}) is required to obtain good models.

Unfortunately there are numerous examples of QSARs/QSPRs having been developed on training sets with a narrow range of endpoint values. A QSAR for toxicity of 25 antibiotics had a log $1/\mathrm{LD}_{50}$ range of 0.84 log units [73], and a QSPR for the melting point of substituted anilines had a melting point range of 244.5–461.5 K (see [74]). In some cases, of course (as in the latter case cited above), it is not possible to have a sufficiently wide range of endpoint values.

Published datasets sometimes include compounds with weak activities reported, for example, $\mathrm{LD}_{50} > 300\,\mathrm{mg/kg}$; these are termed ‘censored data’. Weakly active compounds are of value in QSAR analysis, and a QSAR practitioner might be tempted to use censored data incorrectly by taking, in the example given above, LD_{50} to be equal to 300 mg/kg, or perhaps some multiple of it. We are not aware of any QSAR publications in which this has been done, but would draw attention to work by Borth and Wilhelm [75,76], who have developed a method for the use of censored data in QSAR analysis.

2.12 *Over-fitting of data*

It is tempting, when developing a QSAR/QSPR, to try to maximise the fit of the data to the model, perhaps by the use of numerous descriptors, by the deletion of outliers or by the use of a particular statistical technique. Topliss and Costello [77] showed that, in order to minimise the risk of chance correlations, the ratio of number of training set compounds to the number of descriptors in the QSAR/QSPR should be at least 5 : 1. There are many instances in the literature in which this ‘rule’ has been broken, with perhaps the worst example being that in which 9 descriptors were used to model the aquatic toxicities of 12 alcohols [78]. Another example is the use of 5 descriptors to model the mutagenicity of 17 compounds [79].

Topliss and Edwards [80] drew attention to the risk of chance correlations not merely by the use of large numbers of descriptors in a QSAR/QSPR relative to the number of training set compounds, but also by using a large pool of descriptors from which to select the final descriptors to be used in the model. They showed, for example, that if a pool of 70 descriptors were used, a minimum of about 98 training set compounds would be needed to yield $r^2 \geq 0.8$ with a chance correlation level of less than 0.01. With the modern tendency to use pools of several hundred descriptors, one could imagine that the risk of chance correlations would be extremely high. Our experience and that of others [31] is that the risks of chance correlations is much lower than the Topliss and Edwards work suggests, but we suggest that, as a safeguard, *y*-scrambling should always be carried out on the developed QSAR/QSPR. Strictly, the scrambling procedure should be performed on the whole descriptor pool, and not just on the descriptors already selected for the QSAR/QSPR [31,81].

2.13 *Use of excessive numbers of descriptors in a QSPR*

Aptula et al. [56] have pointed out that QSARs/QSPRs containing large numbers of descriptors are difficult to interpret, because of their complexity. The principle of Occam's razor (sometimes called the principle of parsimony) applies here: 'One should not increase beyond what is necessary the number of entities required to explain anything'. We suggest that five or six descriptors are generally the maximum that one should generally use in a QSAR/QSPR, partly because it is difficult to comprehend the mechanistic significance of large numbers of descriptors. There are, however, numerous reports of large numbers of descriptors being used; for example 55 descriptors were used to model the aqueous solubility of 1050 compounds [82]. It should be noted that group contributions (and perhaps also electrotopological state indices, since they are in effect a special type of group contribution) are exempt from Occam's razor, since they are simply counts of the presence of defined sub-structural molecular features.

2.14 *Lack of/inadequate statistics*

Occasionally, even now, QSARs and QSPRs appear in print without statistics, such as that for the prediction of aqueous solubility [83]. A more common occurrence is for some statistical parameters not to be given. Traditionally, the statistical parameters generally reported with a QSAR/QSPR are the square of the correlation coefficient (r^2, or R^2 if more than one descriptor is used), and the standard error of the estimate (s). More recently, the square of the internally cross-validated correlation coefficient (q^2 or Q^2, sometimes written as r^2_{cv} or R^2_{cv}) by the leave-one-out (LOO) procedure, has also been used. It should be noted, however, that internal cross-validation has recently been shown not to be a good indicator of the predictivity of a QSAR/QSPR [10,11,56]. It should also be mentioned here that there appears to be a widespread misconception as to the numerical value of r^2 (or R^2). For a given probability, its value varies with the number of data points used to calculate it [84]. Therefore, a higher value of r^2 or R^2 does not necessarily mean a better correlation, when different sizes of datasets are being compared.

A rarely used parameter is the adjusted square of the correlation coefficient (r^2_{adj} or R^2_{adj}), which adjusts for degrees of freedom, i.e. for the number of descriptors in a QSAR/QSPR. It is valuable because it allows a comparison of R^2 values between

QSARs/QSPRs with different numbers of descriptors. In a QSAR study of the aquatic toxicity of alcohols [78] in which up to eight descriptors were used in addition to log *P*, if R^2_{adj} values had been used instead of R^2 values, it would have been obvious that increasing the number of descriptors did not improve the correlation.

We should like to see r^2_{adj} values used much more widely, and their use should be mandatory for comparative purposes.

It is important to have a measure of the error of prediction for a QSAR/QSPR. The standard error of the estimate is the measure given for most QSAR/QSPR correlations, although other error statistics such as standard deviation, root mean square error and mean absolute error are occasionally used instead. A good QSAR/QSPR should have a prediction error close to the measurement error; that is, the QSAR/QSPR is modelling the data as well as is possible. If the prediction error is considerably greater than the experimental error, the model is clearly inadequate. For example, the experimental error on aqueous solubility for a diverse dataset has been reported [85] to be 0.58 log units, so a QSPR prediction error of 2.4 log units [86] or of 0.99 [87] indicates a poor model. If, on the other hand, the prediction error is much less than the experimental error, the QSAR/QSPR has over-fitted the data [8], as indicated in another QSPR study of aqueous solubility [88] yielding a standard error of 0.08 log units.

It is important also to consider the standard error on the coefficient of each descriptor in a QSAR/QSPR. These are reported only rarely, but they are valuable in that they indicate the likelihood that a descriptor is incorporated by chance. In general, if the standard error on a coefficient is close to or greater than the magnitude of the coefficient of a descriptor, the presence of that descriptor contributes little or nothing to the QSAR/QSPR, and has been selected by chance. For example, in a study of human skin–water partition coefficients [89], a five-descriptor QSPR was reported with one of the descriptors having a coefficient value of 0.024 and a standard error of 0.137; that descriptor should have been discarded.

A more quantitative alternative to the use of standard errors of coefficients is to use the probability (*p*) values associated with each descriptor. Again, these are reported but rarely. The *p*-value is the probability that the descriptor is there by chance, and a *p*-value of $p \leq 0.05$ (a probability of $\leq 5\%$) is generally taken as acceptable, although it should be noted that *p*-values have little meaning if the descriptors are selected from a large pool. The results shown in Table 4, which are taken from an actual (unpublished) QSAR analysis of cognition-enhancing drugs, exemplify this; the dependent variable was the logarithm of the latent time in the rat scopolamine-induced passive avoidance reflex amnesia model. The QSAR statistics were $n = 16$, $r^2 = 0.808$, $Q^2 = 0.665$, $s = 0.183$.

Clearly, the *p*-value for the maximum negative charge descriptor is unacceptably high, and this is confirmed by the fact that the standard error of the coefficient is almost as high as the coefficient value. This descriptor was therefore discarded from the correlation. Note that it is generally considered that the above guidelines do not apply to the constant term.

We recommend that either standard errors of coefficients or *p*-values be reported for all published QSARs/QSPRs.

The Fisher statistic or variance ratio (*F*) is now widely used as a QSAR/QSPR statistic. It gives an indication of the fit of a regression equation to the training set data, and is defined as the ratio of the explained mean square to the residual mean square. Tables of *F* statistics are available in statistics textbooks, and values are listed for different confidence levels such as 0.05 (i.e. a 5% probability that the correlation has occurred by chance).

Table 4. Standard errors and p-values of coefficients in QSAR model of 16 cognition-enhancing drug candidates.

Descriptor	*Coefficient*	*Standard error of coefficient*	*p-value*
Number of double bonds	35.9	4.9	<0.001
Number of halogen atoms	2.68	0.64	0.001
Maximum negative charge	29.8	24.2	0.241
Constant	−8.5	4.7	0.091

Dearden and Cronin [90] have given an example of their use: Equation (7) represents the modelling of a biological response with log P and $(\log P)^2$:

$$\log 1/C = 3.416 \log P - 0.942(\log P)^2 + 5.816 \tag{7}$$

$$n = 17, \quad r^2 = 0.880, \quad s = 0.141, \quad F = 11.9$$

An F statistics table for a given confidence level will have columns for the number of independent variables or descriptors (m) and rows for the number of degrees of freedom $(n-m-1)$; the 1 accounts for the constant term in the equation. The number of degrees of freedom for Equation (7) are $(17-2-1)=14$. From the appropriate F statistics table, the F statistic for $m=2$ and $(n-m-1)=14$, for the 0.001 confidence level, is 11.78. As the F statistic for Equation (7) is slightly greater than this, the equation is valid at the 0.001 confidence level; that is, there is only a 0.1% probability that Equation (7) is a chance correlation. Strictly, this should be reported as $F_{2,14}=11.9$; $F_{2,14}$, α, $0.001=11.78$. However, most authors simply report the F statistic and leave the reader to look up the relevant statistical table to determine the confidence level. We recommend that full F statistic information is given for each QSAR/QSPR, to show the confidence level and to save readers having to look up confidence levels for themselves.

In summary, we recommend that the following statistical information is given for each published QSAR/QSPR: n, r^2 or R^2, q^2 or Q^2, R^2_{adj}, s, and full F statistics (including p-values).

It is pertinent here also to comment on the different measures of error that are used in QSAR/QSPR studies. The commonest is standard error of the estimate (s), but others are standard deviation, root mean square error and mean absolute error. It would be helpful if QSAR/QSPR practitioners could standardise on one of these; we suggest standard error, since that is the most widely used.

2.15 *Incorrect calculation*

It is usually assumed, by editors and manuscript reviewers, that the authors of a paper have made their calculations correctly. Indeed, in many cases it is impossible to check the authors' calculations because insufficient data are given. However, incorrectly calculated QSARs and QSPRs have certainly been published, which is rather worrying, because we have no idea how widespread this problem may be. In a study of the glycine conjugation rate in liver of 23 benzoic acid derivatives [91], a five-descriptor QSAR was reported with $r^2=0.958$, $s=0.14$ and $F=79$. As all descriptor values were given we were able to check the authors' calculations, and found the statistics to be $r^2 = 0.298$, $s = 0.56$, and $F = 1.4$. We also found that we could not reproduce the statistics given for a prediction set QSPR reported for the aqueous solubility of 40 organic chemicals [83].

2.16 *Lack of descriptor auto-scaling*

Descriptors usually cover different numerical ranges. For example, Dearden et al. [53] developed a QSAR for the toxicity of 2- and 3-substituted nitrobenzenes to the aquatic ciliate *Tetrahymena pyriformis* using three descriptors: $\log D$, the Swain–Lupton F value, and $|dC_{\text{ox}}|$, the modulus of change of charge on the nitro oxygen atom. The $\log D$ values ranged from −3.572 to +3.773, the F values from −0.04 to +0.67, and the $|dC_{\text{ox}}|$ values from 0.0000 to 0.0646. This can give rise to two problems. First, it is very difficult to assess the relative contributions of each descriptor to the model, and second the descriptor(s) with large numerical values would dominate those with small numerical values [7], so that the validity of the statistical analysis could be compromised.

The difficulties can be overcome by the use of auto-scaling, in which the mean is subtracted from the descriptors and the resultant values divided by the standard deviation. Auto-scaled descriptors have a mean of zero and a variance of one, and are less susceptible to the influence of compounds with extreme values [3]. A good example of descriptors that are (approximately) auto-scaled is the so-called Abraham descriptors. Abraham et al. [92] developed the following QSAR for inhalation anaesthesia:

$$\log 1/\text{MAC} = -0.781 - 0.071\,\text{MR}_{\text{ex}} + 1.548\,\text{Pol} + 3.684\,\text{HA} + 1.372\,\text{HB} + 0.697\,\text{L} \quad (8)$$

$$n = 74, \quad r^2 = 0.985, \quad s = 0.198, \quad F = 906.2$$

where MAC is the alveolar concentration that prevents movement in 50% of subjects, MR_{ex} is the excess molar refractivity (a measure of polarisability), Pol is the polarity, HA is the hydrogen bond donor acidity, HB is the hydrogen bond acceptor basicity, and L is the logarithm of the gas–hexadecane partition coefficient (a measure of molecular size). As the descriptor values are approximately auto-scaled, it can readily be seen that the most important contribution to anaesthesia is hydrogen bond donor ability, whilst the least important is polarisability. In fact the polarisability term is statistically insignificant, since its standard error of 0.094 is greater than the value of its coefficient.

We recommend that auto-scaling be carried out on descriptors before development of a QSAR/QSPR. It may be noted that some commercial software programs, such as SIMCA [93], have an auto-scaling facility.

2.17 *Misuse/misinterpretation of statistics*

Statistics is not an easy subject, and most QSAR/QSPR practitioners are not statisticians. It is therefore relatively easy to misuse or misinterpret QSAR/QSPR statistics. There are very few books dealing with QSAR/QSPR statistics; a good one is that by Livingstone [3], which is currently being updated. Devillers and Doré [94] have given a survey of some of the statistics available on the Internet for QSAR/QSPR. Two very useful sites are those of the Scripps Institute [95] and QSAR World [96].

The study by Romanelli et al. [78] has already been referred to. Their initial correlation of the aquatic toxicity of alcohols with log P yielded $r^2 = 0.9930$, $s = 0.0246$, values which are surprisingly good when one considers that there is a degree of experimental error in the toxicity data. They then introduced up to eight additional descriptors in an attempt to 'improve' the correlation; this was unacceptable, (a) because the correlation was already at least as good as one could expect, (b) because their 'improvement' broke the Topliss and Costello rule (see Section 2.12) and (c) because they ignored the principle of Occam's razor (see Section 2.13).

A QSPR study of aqueous solubility [88] used fuzzy ARTMAP statistics to obtain a standard error of 0.08 log units. Since the standard error of measurement of aqueous solubility is about 0.6 log units [38,85], there was misuse or misinterpretation of statistics, for the error of prediction cannot, if valid, be significantly less than the error of measurement [3].

2.18 *No consideration of distribution of residuals*

There are two types of error that contribute to model predictions, namely random error and systematic error. The former is an indication of the irreproducibility in the data and/or the descriptor values, whilst the latter usually result from biases in measurement or calculation, and could indicate poor selection of descriptors.

It is not essential to check for error distribution in a QSAR/QSPR model, but added confidence is given to the model if lack of systematic error can be demonstrated. A simple plot of residuals (i.e. measured minus predicted values of the property studied) against measured values will be random around the (residual = 0) line. If, however, there is systematic error all or most of the residuals may be on one side of the (residual = 0) line, and may also show a regular variation of residuals with increasing measured values. A recent QSAR study of apoptosis [97] claimed that a residuals plot showed no systematic error. However, re-analysis of their data showed a significant non-zero gradient of the residuals plot, suggesting a systematic contribution to the error. Fang et al. [98] have recommended the use of Kriging models to improve performance.

Abraham et al. [99] have pointed out that the same information can be obtained by comparing average error (AE; i.e. taking account of the sign of an error) with average absolute error (AAE; i.e. ignoring the sign of an error). If the AE values are very small, there is no systematic error, and the AAE values represent random error. If, however, AE and AAE values are identical or very similar, there is systematic error.

It would be helpful to see residual plots included in QSAR/QSPR publications.

2.19 *Inadequate training/test set selection*

Ideally, training set data should be well distributed across the range of endpoint values. This is not always possible in practice [3], but very poor distribution should be avoided [100]. The extreme case is two widely separated clusters of data points through which a straight line has been fitted, examples of which are two graphs depicting the correlation of observed and predicted toxicities of aldehydes and amines to *Tetrahymena pyriformis* [101]. A less obvious case concerns a QSPR study of chemical reactivity of acrylates and methacrylates [102]; a good r^2 value was obtained, but when we plotted the results graphically, it was clear that the correlation resulted from an acrylate cluster and a methacrylate cluster. A related problem arises when just one datum point is far from the others, and thus heavily influences the correlation.

In general, datasets are divided into training and test sets either randomly, or by an activity-range algorithm [103]. However, it has been shown [104–106] that with these approaches, predictive models are not obtained in most cases. Various methods, including clustering [105,107], D-optimal experimental design and Kohonen artificial neural networks [39] and others [108,109], have been proposed for rational training/test set selection, and have been shown to yield more predictive QSARs/QSPRs. Schultz et al. [110] stated that 'whilst selection of the training set of chemicals for a QSAR should be based on diversity, we feel that selection of the chemicals for validation should be based on representivity'. Whatever the best approach may be, the selection of training and testing sets can have a substantial influence on a model's statistics and performance. It would appear that more research is desirable into this area as the judicious selection of training and testing sets can provide for a wide range of model statistics [105].

It is recommended that a rational approach to training and test set selection be adopted in all QSAR/QSPR analyses.

2.20 *Failure to validate a QSPR correctly*

It is clear from section 2.19 that a key stage in QSAR/QSPR statistical validation is rational training and test set selection, which is very rarely done at present. In addition, as pointed out in the introduction, it is now generally accepted that external validation, using data not included in the training set, is the only way to ensure good predictivity of a QSAR/QSPR [6,10,11,39,111]. It may be noted that at least four journals (*SAR and QSAR in Environmental Research*, *QSAR and Combinatorial Science*, *Journal of Medicinal Chemistry* and *Journal of Chemical Information and Modeling*) require external validation of published QSARs/QSPRs, and the last two also have additional requirements [112].

Tropsha et al. [81] discussed a number of validation strategies, including randomisation of the modelled property (*y*-scrambling), multiple leave-many-out cross-validations, bootstrapping and external validation using rational division of a dataset into training and test sets. Their recommendation is first to carry out internal validation, and then to undertake external validation. Ideally, they say, the procedure of training and test set selection and external validation should be repeated several times so as to identify the QSAR/QSPR model for the smallest training set that affords adequate prediction power for the largest test set.

Tropsha et al. [81] also say that the ideal splitting of a dataset leads to a test set such that each of its members is close to at least one point of the training set. However, whilst such a strategy will almost certainly lead to good prediction of test set activities (unless 'activity cliffs' are present [49,50]), Tetko [113] has suggested that such supervised selection of a test set so as to fall within the domain of applicability of the training set does not measure the true predictive ability of the QSAR/QSPR. We concur with Tetko [113] that selecting a test set so that each member is close to at least one member of the training set could lead to artificially high predictivity. On the other hand, some test set selection is essential, otherwise poor predictivity will certainly result. Devillers et al. [114] recognised this problem many years ago, and proposed the use of two test sets: an in-sample set containing compounds widely represented in the training set and an out-of-sample set containing compounds only weakly represented in the training set. This is an eminently sensible approach, but unfortunately has not been widely adopted. Tetko et al. [115] have suggested an alternative strategy of five-fold cross-validation with variable selection, which performs a blind external prediction of the validation subset in each validation fold. There will undoubtedly be further discussion, if not controversy, on this subject.

2.21 *Lack of mechanistic interpretation*

The guidance for the OECD Principles for the Validation of (Q)SARs [13] states that 'It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR (Principle 5), or that there even be multiple mechanistic interpretations of a given model. The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intent of Principle 5 is not to reject models that have no apparent mechanistic basis, but to ensure that some consideration is given to the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted, and to ensure that this association is documented'.

The OECD report on the principles for the validation of (Q)SARs/(Q)SPRs [14] suggests that the following questions be asked regarding the mechanistic basis of a QSAR/QSPR.

(1) Do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?
(2) Can any literature references be cited in support of the purported mechanistic basis of the QSAR/QSPR?

If the answers to both questions are affirmative, one may have a degree of confidence in the proposed mechanism of action. If the answer to one or both questions is negative, then of course the level of confidence will be lower. In either case, it must be borne in mind that the existence of a correlation does not imply causality. As mentioned earlier, a study of the toxicity of a series of alkyl ethers to the mouse [28] yielded a good correlation with first-order molecular connectivity, an indicator of molecular size, whereas Cronin and Schultz [6] pointed out that a very similar correlation obtains with partition coefficient. It is known that partition-controlled absorption is much more common than is molecular size-controlled absorption [116], so it would be reasonable in this case to postulate a mechanism based on partitioning. Nevertheless, great care must be taken when trying to interpret QSARs/QSPRs mechanistically. Quite apart from the risk of chance correlations mentioned in Sections 2.4 and 2.12, the existence of confounding factors should be considered. In the famous and oft-quoted example of a correlation between the number of human births and the stork population in the Netherlands, the confounding factor could be the number of houses and/or chimneys: as the population grows, more houses (with chimneys) are built, and so more nesting sites for storks are available.

Rowe [117] gives an intriguing correlation, using data from the UK Office of National Statistics, between annual deaths from liver disease and the proportion of households owning a microwave oven during the period 1991–2001. The correlation is a good one ($r^2 = 0.91$, $p = 0.001$), and one could be tempted to suggest that the correlations indicate that microwaves have an adverse effect on liver cells. However, there are almost certainly confounding factors, for example increasing standards of living that led people both to buy more microwave ovens and to drink more alcohol.

Johnson [118] has pointed out that 'QSAR has devolved into a perfectly practiced art of logical fallacy; *cum hoc ergo propter hoc* (with this, therefore because of this)'. He also comments that a statement to the effect that, say, a particular descriptor represents hydrogen bonding is not a mechanistic interpretation.

In addition, Johnson [118] has stated that 'rarely, if ever, are any designed experiments presented to test or challenge the interpretation of the descriptors... Statistical methodologies should be a tool of QSAR but instead have often replaced the craftsman tools of our trade – rational thought, controlled experiments, and personal observation'.

It is recommended that very careful consideration be given to mechanistic interpretations of QSARs/QSPRs, and that referees should not be unduly critical of QSAR/QSPR manuscripts that do not offer mechanistic interpretations.

3. Conclusions

Many published QSAR/QSPR analyses still contain errors, despite there being numerous published recommendations and a published set of five requirements (the OECD Principles for the Validation of (Q)SARs) for the avoidance of error. We have identified 21 types of error that are found throughout the QSAR/QSPR literature up to the present, and each of these is allocated to the relevant OECD principle. Published examples of each type of error have been given, including several from our own publications. Most of the errors relate to principle 2 (an unambiguous algorithm), principle 3 (a defined domain of applicability) and particularly to principle 4 (appropriate measures of goodness of fit, robustness and predictivity), which relates to statistical aspects of QSAR/QSPR modelling. Where appropriate, we have made recommendations for the avoidance of such errors, in order to help researchers, authors and referees. We hope that our investigations and recommendations will improve the validity and usefulness of future QSAR/QSPR studies.

References

[1] C. Hansch, P.P. Maloney, T. Fujita, and R.M. Muir, *Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients*, Nature 194 (1962), pp. 178–180.

[2] E.J. Mills, *On melting point and boiling point as related to composition*, Phil Mag. Ser. 5 17 (1884), pp. 173–187.

[3] D. Livingstone, *Data Analysis for Chemists*, Oxford University Press, Oxford, UK, 1995.

[4] J.D. Walker, J.C. Dearden, T.W. Schultz, J. Jaworska, and M.H.I. Comber, *QSARs for new practitioners*, in *QSARs for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications*, J.D. Walker, ed., SETAC Press, Pensacola, FL, 2003, pp. 3–18.

[5] J.D. Walker, J. Jaworska, M.H.I. Comber, T.W. Schultz, and J.C. Dearden, *Guidelines for developing and using quantitative structure–activity relationships*, Environ. Toxicol. Chem. 22 (2003), pp. 1653–1665.

[6] M.T.D. Cronin and T.W. Schultz, *Pitfalls in QSAR*, J. Theoret. Chem. (Theochem) 622 (2003), pp. 39–51.

[7] L. Eriksson, J. Jaworska, A.P. Worth, M.T.D. Cronin, R.M. McDowell, and P. Gramatica, *Methods for reliability and uncertainty assessment for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Persp. 111 (2003), pp. 1361–1375.

[8] D.J. Livingstone, *Building QSAR models: a practical guide*, in *Predicting Chemical Toxicity and Fate*, M.T.D. Cronin and D.J. Livingstone, eds., CRC Press, Boca Raton, FL, 2004, pp. 151–170.

[9] A. Tropsha and A. Golbraikh, *Predictive QSAR modeling workflow, model applicability domains, and virtual screening*, Curr. Pharmaceutical Design 13 (2007), pp. 3494–3504.

[10] A. Golbraikh and A. Tropsha, *Beware of q^2!*, J. Mol. Graph. Modell. 20 (2002), pp. 269–276.

[11] P. Gramatica, *Principles of QSAR models validation: internal and external*, QSAR Comb. Sci. 26 (2007), pp. 694–701.

[12] J.S. Jaworska, M. Comber, C. Auer, and C.J. Van Leeuwen, *Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints*, Environ. Health Persp. 111 (2003), pp. 1358–1360.

[13] OECD Principles for the Validation of (Q)SARs, http://www.oecd.org/dataoecd/33/37/37849783.pdf (last accessed November 2008).

[14] OECD, Environment Directorate, *Joint Meeting of The Chemicals Committee and The Working Party on Chemicals, Pesticides and Biotechnology*, http://www.olis.oecd.org/olis/2004doc.nsf/LinkTo/NT00009192/$FILE/JT00176183.PDF (last accessed November 2008).

[15] S.H. Unger and C. Hansch, *On model building in structure–activity relationships. A reexamination of adrenergic blocking activity of β-halo-β-arylalkylamines*, J. Med. Chem. 16 (1973), pp. 745–749.

[16] G.I. Poda, D.P. Landsittel, K. Brumbaugh, D.S. Sharp, H.F. Frasch, and E. Demchuk, *Random sampling or 'random' model in skin flux measurements? [Commentary on "Investigation of the mechanism of flux across human skin in vitro by quantitative structure-permeability relationships"]*, Eur. J. Pharm. Sci. 14 (2001), pp. 197–200.

[17] K. Yoshida and T. Niwa, *Quantitative structure–activity studies on inhibition of HERG potassium channels*, J. Chem. Inf. Model. 46 (2006), pp. 1371–1378.

[18] A. Coi, I. Massarelli, L. Murgia, M. Saraceno, V. Calderone, and A.M. Bianucci, *Prediction of hERG potassium channel affinity by the CODESSA approach*, Bioorg. Med. Chem. 14 (2006), pp. 3153–3159.

[19] S.R. Johnson and P.C. Jurs, *Prediction of acute mammalian toxicity from molecular structure for a diverse set of substituted anilines using regression analysis and computational neural networks*, in *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*, H. van de Waterbeemd, B. Testa, and G. Folkers, eds., Wiley-VCH, Weinheim, 1997, pp. 29–48.

[20] CDC, Registry of Toxic Effects of Chemical Substances (RTECS) http://www.cdc.gov/niosh/rtecs/ (last accessed November 2008).

[21] E.J. Matthews, N.L. Kruhlak, R.D. Benz, and J.F. Contrera, *Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data*, Curr. Drug Discov. Technol. 1 (2004), pp. 61–76.

[22] EPA, FDAMDD: FDA Maximum (Recommended) Daily Dose, http://www.epa.gov/ncct/dsstox/sdf_fdamdd.html (last accessed November 2008).

[23] A. Geronikaki, P. Vicini, G. Theophilidis, A. Lagunin, V. Poroikov, and J.C. Dearden, *Study of local anesthetic activity of some derivatives of 3-aminobenzo-[d]-isothiazole*, SAR QSAR Environ. Res. 14 (2003), pp. 485–495.

[24] J.C. Dearden, S.J.A. Bradburne, M.T.D. Cronin, and P. Solanki, *The physical significance of molecular connectivity*, in *QSAR 1988*, J.E. Turner, M.W. England, T.W. Schultz, and N.J. Kwaak, eds., US Department of Energy, Oak Ridge, TN, 1988, pp. 43–50.

[25] X. Lu, S. Tao, J. Cao, and R.W. Dawson, *Prediction of fish bioconcentration factors of nonpolar organic pollutants based on molecular connectivity indices*, Chemosphere 39 (1999), pp. 987–999.

[26] M. Nendza, *Structure–Activity Relationships in Environmental Sciences*, Chapman & Hall, London, 1998.

[27] D.A. Basketter, D.W. Roberts, M. Cronin, and E.W. Scholes, *The value of the local lymph-node assay in quantitative structure-activity investigations*, Contact Dermatitis 27 (1992), pp. 137–142.

[28] M. Randić and S.C. Basak, *On use of the variable connectivity index $^1\chi^f$ in QSAR: toxicity of aliphatic ethers*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 614–618.

[29] L. Eriksson, J.L.M. Hermens, E. Johansson, H.J.M. Verhaar, and S. Wold, *Multivariate analysis of aquatic toxicity data with PLS*, Aquat. Sci. 57 (1995), pp. 217–241.

[30] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.

[31] H. Kubinyi, *Validation and predictivity of QSAR models*, in *QSAR and Molecular Modelling in Rational Design of Bioactive Molecules*, E.A. Sener and I. Yalcin, eds., CADDD Society, Ankara, 2006, pp. 30–33.

[32] W. Tong, R. Perkins, R. Strelitz, E.R. Collantes, S. Keenan, W.J. Welsh, W.S. Branham, and D.M. Sheehan, *Quantitative structure–activity relationships (QSARs) for estrogen binding to the estrogen receptor: predictions across species*, Environ. Health Persp. 106 (1997), pp. 1116–1124.

[33] O. Ivanciuc, *QSAR comparative study of Wiener descriptors for weighted molecular graphs*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 1412–1422.

[34] P. Gramatica, N. Navas, and R. Todeschini, *3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs)*, Chemomet. Intell. Lab. Sys. 40 (1998), pp. 53–63.

[35] H. Niska, K. Tuppurainen, J.-P. Skön, A.K. Mallett, and M. Kolehmainen, *Characterisation of the chemical and biological properties of molecules with QSAR/QSPR and chemical grouping, and its application to a group of alkyl ethers*, SAR QSAR Environ. Res. 19 (2008), pp. 263–284.

[36] J.C. Dearden, *In silico prediction of aqueous solubility*, Expert Opin. Drug. Discov. 1 (2006), pp. 31–52.

[37] E. Rytting, K.A. Lentz, X.-Q. Chen, F. Qian, and S. Venkatesh, *Aqueous and cosolvent solubility data for drug-like organic compounds*, Am. Assoc. Pharm. Sci. J. 7 (2005), pp. E78–E105.

[38] W.L. Jorgensen and E.M. Duffy, *Prediction of drug solubility from structure*, Adv. Drug. Deliv. Rev. 54 (2002), pp. 355–366.

[39] P. Gramatica, P. Pilutti, and E. Papa, *Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modelling*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 1794–1802.

[40] T. Abshear, G.M. Banik, M.L. D'Souza, K. Nedwed, and C. Peng, *A model validation and consensus building environment*, SAR QSAR Environ. Res. 17 (2006), pp. 311–321.

[41] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Öberg, P. Dao, A. Cherkasov, and I.V. Tetko, *Combinatorial QSAR modeling of chemical toxicants against Tetrahymena pyriformis*, J. Chem. Inf. Model. 48 (2008), pp. 766–784.

[42] T. Ghafourian and J.C. Dearden, *The use of atomic charges and orbital energies as hydrogen-bonding-donor parameters for QSAR studies: comparison of MNDO, AM1 and PM3 methods*, J. Pharm. Pharmacol. 52 (2000), pp. 603–610.

[43] J.C. Dearden, T.I. Netzeva, and R. Bibby, *A comparison of commercially available software for the prediction of partition coefficient*, in *Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, M. Ford, D. Livingstone, J. Dearden, and H. van de Waterbeemd, eds., Blackwell, Oxford, 2003, pp. 169–170.

[44] J.C. Dearden and G. Schüürmann, *Quantitative structure–property relationships for predicting Henry's law constant from molecular structure*, Environ. Toxicol. Chem. 22 (2003), pp. 1755–1770.

[45] T. Hartung, S. Bremer, S. Casati, S. Coecke, R. Corvi, S. Fortaner, L. Gribaldo, M. Halder, S. Hoffmann, A. Janusch Roi, P. Prieto, E. Sabbioni, L. Scott, A. Worth, and V. Zuang, *A modular approach to the ECVAM principles on test validity*, ATLA 32 (2004), pp. 467–472.

[46] A. Roncaglioni, *Data exploration and knowledge extraction: their application to the study of endocrine disrupting chemicals*, Ph.D. Thesis, The Open University, Milton Keynes, UK, 2008.

[47] T.I. Netzeva, A. Worth, T. Aldenberg, R. Benigni, M.T. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D. Roberts, T. Schultz, D.W. Stanton, J.J. van de Sandt, W. Tong, G. Veith, and

C. Yang, *Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. The report and recommendations of ECVAM Workshop 52*, ATLA 33 (2005), pp. 155–173.

[48] J. Jaworska, N. Nikolova-Jeliaskova, and T. Aldenberg, *QSAR applicability domain estimation by projection of the training set in descriptor space: a review*, ATLA 33 (2005), pp. 445–459.

[49] H. Kubinyi, *Chemical similarity and biological activities*, J. Braz. Chem. Soc. 13 (2002), pp. 717–726.

[50] G.M. Maggiora, *On outliers and activity cliffs – why QSAR often disappoints*, J. Chem. Inf. Model. 46 (2006), p. 1535.

[51] E. Papa, F. Villa, and P. Gramatica, *Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow)*, J. Chem. Inf. Model. 45 (2005), pp. 1256–1266.

[52] TOPKAT software, http://www.accelrys.com (last accessed November 2008).

[53] J.C. Dearden, M.T.D. Cronin, T.W. Schultz, and D.T. Lin, *QSAR study of the toxicity of nitrobenzenes to Tetrahymena pyriformis*, Quant. Struct.–Act. Relat. 14 (1995), pp. 427–432.

[54] A. Niazi, S. Jameh-Bozorghi, and D. Nori-Shargh, *Prediction of toxicity of nitrobenzenes using ab initio and least squares support vector machines*, J. Hazardous Mat. 151 (2008), pp. 603–609.

[55] R.L. Lipnick, *Outliers: their origin and use in the classification of molecular mechanisms of toxicity*, Sci. Tot. Environ. 109/110 (1991), pp. 131–153.

[56] A.O. Aptula, N.G. Jeliazkova, T.W. Schultz, and M.T.D. Cronin, *The better predictive model: high q^2 for the training set or low root mean square error of prediction for the test set?*, QSAR Comb. Sci. 24 (2005), pp. 385–396.

[57] S. Doniger, T. Hofmann, and J. Yeh, *Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms*, J. Comput. Biol. 9 (2002), pp. 849–864.

[58] N.T. Hansen, I. Kouskoumvekaki, F.S. Jørgensen, S. Brunak, and S.Ó. Jónsdóttir, *Prediction of pH-dependent aqueous solubility of druglike molecules*, J. Chem. Inf. Model. 46 (2006), pp. 2601–2609.

[59] D. Young, T. Martin, R. Venkatapathy, and P. Harten, *Are the chemical structures in your QSAR correct?*, QSAR Comb. Sci. 27 (2008), pp. 1337–1345.

[60] G.L. Flynn, *Physicochemical determinants of skin absorption*, in *Principles of Route-to-Route Extrapolation for Risk Assessment*, T.R. Gerrity and C.J. Henry, eds., Elsevier, Amsterdam, 1990, pp. 93–127.

[61] L.A. Kirschner, R.P. Moody, E. Doyle, R. Bose, J. Jeffery, and I. Chu, *The prediction of skin permeability by using physicochemical data*, ATLA 25 (1997), pp. 359–370.

[62] M.T.D. Cronin, J.C. Dearden, G.P. Moss, and G. Murray-Dickson, *Investigation of the mechanism of flux across human skin in vitro by quantitative structure–permeability relationships*, Eur. J. Pharm. Sci. 7 (1999), pp. 325–330.

[63] H.F. Frasch and D.P. Landsittel, *Regarding the sources of data analyzed with quantitative structure–skin permeability relationship methods (commentary on 'Investigation of the mechanism of flux across human skin in vitro by quantitative structure-permeability relationships')*, Eur. J. Pharm. Sci. 15 (2002), pp. 399–403.

[64] M. Hewitt, J.C. Madden, P.H. Rowe, and M.T.D. Cronin, *Structure-based modelling in reproductive toxicology: (Q)SARs for the placental barrier*, SAR QSAR Environ. Res. 18 (2007), pp. 57–76.

[65] M. Jalali-Heravi, P. Shahbazikhah, B. Zekavat, and M.S. Ardejani, *Principal component analysis-ranking method for the simulation of ^{13}C nuclear magnetic resonance spectra of xanthones using artificial neural networks*, QSAR Comb. Sci. 26 (2007), pp. 764–772.

[66] Cambridgesoft, *ChemBioFinder.com Scientific Database Gateway*, http://chembiofinder.cambridgesoft.com/chembiofinder/SimpleSearch.aspx (last accessed November 2008).

[67] US National Library of Medicine, *ChemIDplus Advanced*, http://chem.sis.nlm.nih.gov/chemidplus (last accessed November 2008).

[68] Daylight Chemical Information Systems Inc., SMILES – A Simplified Chemical Language, http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (last accessed November 2008).

[69] IUPAC, The IUPAC International Chemical Identifier (InChI^{TM}), http://www.iupac.org/inchi (last accessed November 2008).

[70] D.R. Lide, ed., *CRC Handbook of Chemistry and Physics*, 73rd ed., CRC Press, Boca Raton, FL, 1992–1993.

[71] E.J. Delgado, J.B. Alderete, and G.A. Jaña, *A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 1928–1932.

[72] P. Gedeck, B. Rohde, and C. Bartels, *QSAR – how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets*, J. Chem. Inf. Model. 46 (2006), pp. 1924–1936.

[73] F.R. Quinn, *The quantitative structure–activity relationship of rifamycin B amide and hydrazide toxicity in mice*, Il Pharmaco 37 (1982), pp. 3–12.

[74] J.C. Dearden, *The QSAR prediction of melting point, a property of environmental relevance*, Sci. Tot. Environ. 109/110 (1991), pp. 59–68.

[75] D.M. Borth, *Optimal experimental designs for (possibly) censored data*, Chemomet. Intell. Lab. Sys. 32 (1996), pp. 25–35.

[76] D.M. Borth and M.S. Wilhelm, *Confidence limits for normal type I censored regression*, Chemomet. Intell. Lab. Sys. 63 (2002), pp. 117–128.

[77] J.G. Topliss and R.J. Costello, *Chance correlations in structure–activity studies using multiple regression analysis*, J. Med. Chem. 15 (1972), pp. 1066–1068.

[78] G.P. Romanelli, L.F.R. Cafferata, and E.A. Castro, *An improved QSAR study of toxicity of saturated alcohols*, J. Mol. Struct. – Theochem 504 (2000), pp. 261–265.

[79] R. Benigni, L. Conti, R. Crebelli, A. Rodomonte, and M.R. Vari, *Simple and α,β-unsaturated aldehydes: correct prediction of genotoxic activity through structure–activity relationship models*, Environ. Mol. Mutagen. 46 (2005), pp. 268–280.

[80] J.G. Topliss and R.P. Edwards, *Chance factors in studies of quantitative structure–activity relationships*, J. Med. Chem. 22 (1979), pp. 1238–1244.

[81] A. Tropsha, P. Gramatica, and V.K. Gombar, *The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.

[82] D. Erös, G. Kéri, I. Kövesdi, C. Szánti-Kis, G. Mészáros, and L. Örfi, *Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods*, Mini-Rev. Med. Chem. 4 (2004), pp. 167–177.

[83] J. Ghasemi and S. Saaidpour, *QSPR prediction of aqueous solubility of drug-like organic compounds*, Chem. Pharm. Bull. 55 (2007), pp. 669–674.

[84] P.P. Roy, J.T. Leonard, and K. Roy, *Exploring the impact of size of training sets for the development of predictive QSAR models*, Chemomet. Intell. Lab. Sys. 90 (2008), pp. 31–42.

[85] A.R. Katritzky, Y. Wang, S. Sild, T. Tamm, and M. Karelson, *QSPR studies on vapor pressure, aqueous solubility, and the prediction of water–air partition coefficients*, J. Chem. Inf. Comput. Sci. 38 (1998), pp. 720–725.

[86] P. Labute, *A widely applicable set of descriptors*, J. Mol. Graph. Model. 18 (2000), pp. 464–477.

[87] C. Zhong and Q. Hu, *Estimation of the aqueous solubility of organic compounds using molecular connectivity indices*, J. Pharm. Sci. 92 (2003), pp. 2284–2294.

[88] D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, and F. Giralt, *A fuzzy ARTMAP based on quantitative structure-property relationships (QSPRs) for predicting aqueous solubility of organic compounds*, J. Chem. Inf. Comput. Sci. 41 (2001), pp. 1177–1207.

[89] M.H. Abraham and F. Martins, *Human skin permeation and partition: general linear free-energy relationship analyses*, J. Pharm. Sci. 93 (2004), pp. 1508–1523.

[90] J.C. Dearden and M.T.D. Cronin, *Quantitative structure-activity relationships (QSAR) in drug design*, in *Smith and Williams' Introduction to the Principles of Drug Design and Action*, 4th ed., H.J. Smith, ed., Taylor & Francis, Boca Raton, FL, 2006, pp. 185–209.

[91] M. Jalali-Heravi and F. Parastar, *Computer modelling of the rate of glycine conjugation of some benzoic acid derivatives: a QSAR study*, Quant. Struct.–Act. Relat. 18 (1999), pp. 134–138.

[92] M.H. Abraham, W.E. Acree, C. Mintz, and S. Payne, *Effect of anesthetic structure on inhalation anesthesia: implications for the mechanism*, J. Pharm. Sci. 97 (2008), pp. 2373–2384.

[93] Umetrics – Experts in multivariate data analysis and design of experiments, http://www.umetrics.com (last accessed November 2008).

[94] J. Devillers and J.C. Doré, *e-Statistics for deriving QSAR models*, SAR QSAR Environ. Res. 13 (2002), pp. 409–416.

[95] QSAR – Working with statistics, http://www.scripps.edu/rc/softwaredocs/msi/cerius45/qsar/working_with_stats.html (last accessed November 2008).

[96] QSAR World – Statistics, http://www.qsarworld.com/statistics.php (last accessed November 2008).

[97] M.H. Fatemi and S. Gharaghani, *A novel QSAR model for prediction of apoptosis-inducing activity of 4-aryl-4-H-chromenes based on support vector machine*, Bioorg. Med. Chem. 15 (2007), pp. 7746–7754.

[98] K.-T. Fang, H. Yin, and Y.-Z. Liang, *New approach by Kriging models to problems in QSAR*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 2106–2113.

[99] M.H. Abraham, R. Sánchez-Moreno, and J.E. Cometto-Muñiz, *A quantitative structure–activity analysis on the relative sensitivity of the olfactory and the nasal trigeminal chemosensory systems*, Chem. Senses 32 (2007), pp. 711–719.

[100] M.S. Tute, *Principles and practice of Hansch analysis: a guide to structure–activity correlation for the medicinal chemist*, in *Advances in Drug Research*, Vol. 6, N.J. Harper and A.B. Simmonds, eds., Academic Press, London, 1971, pp. 1–77.

[101] D.R. Roy, R. Parthasarathi, B. Maiti, V. Subramanian, and P.K. Chattaraj, *Electrophilicity as a possible descriptor for toxicity prediction*, Bioorg. Med. Chem. 13 (2005), pp. 3405–3412.

[102] A.P. Freidig, H.J.M. Verhaar, and J.L.M. Hermens, *Quantitative structure–property relationships for the chemical reactivity of acrylates and methacrylates*, Environ. Toxicol. Chem. 18 (1999), pp. 1133–1139.

[103] O.A. Raevsky, V.Yu. Grigor'ev, E.E. Weber, and J.C. Dearden, *Classification and quantification of the toxicity of chemicals to guppy, fathead minnow and rainbow trout: Part 1. Nonpolar narcosis mode of action*, QSAR Comb. Sci. 27 (2008), pp. 1274–1281.

[104] A. Golbraikh and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*, J. Comput.-Aided Mol. Design 16 (2002), pp. 357–369.

[105] J.T. Leonard and K. Roy, *On selection of training and test sets for the development of predictive QSAR models*, QSAR Comb. Sci. 25 (2006), pp. 235–251.

[106] E. Furusjö, A. Svenson, M. Rahmberg, and M. Andersson, *The importance of outlier detection and training set selection for reliable environmental QSAR predictions*, Chemosphere 63 (2006), pp. 99–108.

[107] L. Eriksson, E. Johansson, M. Müller, and S. Wold, *On the selection of the training set in environmental QSAR analysis when compounds are clustered*, J. Chemom. 14 (2000), pp. 599–616.

[108] A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, and A. Tropsha, *Rational selection of training and test sets for the development of validated QSAR models*, J. Comput.-Aided Mol. Design 17 (2003), pp. 241–253.

[109] B. Hemmateenajad, K. Javadnia, and M. Elyasi, *Quantitative structure–retention relationship for the Kovats retention indices of a large set of terpenes: a combined data splitting–feature selection strategy*, Anal. Chim. Acta 592 (2007), pp. 72–81.

[110] T.W. Schultz, T.I. Netzeva, and M.T.D. Cronin, *Selection of data sets for QSARs: analyses of Tetrahymena toxicity from aromatic compounds*, SAR QSAR Environ. Res. 14 (2003), pp. 59–81.

[111] R. Benigni and C. Bossa, *Predictivity of QSAR*, J. Chem. Inf. Model. 48 (2008), pp. 971–980.

[112] W.L. Jorgensen, *QSAR/QSPR and proprietary data*, J. Chem. Inf. Model. 46 (2006), p. 937.

[113] I.V. Tetko, Personal communication to J.C. Dearden, 16 June 2008.

[114] J. Devillers, S. Bintein, D. Domine, and W. Karcher, *A general QSAR model for predicting the toxicity of organic chemicals to luminescent bacteria (Microtox® test)*, SAR QSAR Environ. Res. 4 (1995), pp. 29–38.

[115] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, and A. Varnek, *Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis; focusing on applicability domain and overfitting by variable selection*, J. Chem. Inf. Model. 48 (2008), pp. 1733–1746.

[116] C. Hansch and A. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, Vol. 1, American Chemical Society, Washington, DC, 1995.

[117] P. Rowe, *Essential Statistics for the Pharmaceutical Sciences*, Wiley, Chichester, 2007.

[118] S.R. Johnson, *The trouble with QSAR (or how I learned to stop worrying and embrace fallacy)*, J. Chem. Inf. Model. 48 (2008), pp. 25–26.