QSAR & Combinatorial Science QCS

The Better Predictive Model: High q^2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set?

Aynur O. Aptula[a], **Nina G. Jeliazkova**[b], **Terry W. Schultz**[c], **and Mark T. D. Cronin**[a]*

[a] School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L33AF, England, e-mail: m.t.cronin@livjm.ac.uk, Tel: +44 151 231 2066, Fax: +44 151 231 2170
[b] Institute of Parallel Processing, Bulgarian Academy of Sciences, 25A "Acad. G. B.onchev" Street, Sofia 1113, Bulgaria
[c] Department of Comparative Medicine, College of Veterinary Medicine, The University of Tennessee, 2407 River Drive, Knoxville, TN 37996-4543, USA

All models are wrong, but some are useful.
George E. P. Box

Keywords: phenol toxicity, model complexity, validation, QSAR, RMSE, q^2

Received: September 9, 2004; Accepted: November 3, 2004

Abstract

The process of validation of computational models (e.g., QSARs) may become the most important step in their development. Different requirements for the reliability and predictability of QSAR models have been described in the literature. Despite these formal recommendations there are few practical rules as to when to cease adding variables to a QSAR (i.e., what is an appropriate level of complexity of the model). In this work the influence of model complexity to statistical fit and error have been investigated using toxicity data for 200 phenols to the ciliated protozoan *Tetrahymena pyriformis* when applying a test set of a further 50 compounds. The results from this investigation showed that some important factors play a role in the definition of a good and reliable QSAR. These include the fact that q^2 is not a good criterion for a model predictivity; that outliers should not necessarily be deleted as this may reduce the chemical space of the model; the number of descriptors in a multivariate model should be chosen carefully to avoid model under- and over-estimation; and that an appropriate number of dimensions is required for PLS modelling.

1 Introduction

Due to the increasing requirement for alternative methods to *in vivo* toxicity testing, a variety of *in vitro* and computational methods are being proposed for the toxicological assessment of chemicals. Within the European Union there is a special emphasis in this area due to the likely demands of the Registration, Evaluation and Authorisation of CHemicals (REACH) legislation [1]. In particular computational models, including quantitative structure-activity relationships (QSARs) may be applied for the assessment of low tonnage chemicals [2, 3].

Computational models for toxicity prediction are among the alternative methods considered to animal testing. They are regularly used by regulatory agencies in the United States of America [4, 5] and their greater use is being contemplated in the European Union [2, 3]. In this context these rapid and low cost methods are attractive to use in prioritising untested chemicals, filling data gaps and for classification and labelling. In order for any alternative method to gain widespread regulatory acceptance (especially within the European Union), it needs to undergo a formal process of validation. The process of validation of an *in vitro* technique is well established [6], and a similar procedure is being developed for computational models [2, 3, 7]. With regard to QSARs the process of validation may become possibly the most important step in model building. Currently, formal validation is also one of the most overlooked steps in model development. For many in the QSAR community, the validation of a model is little more than an assessment of statistical fit, and occasionally predictivity using cross-validation techniques. However, it is now being accepted that validation is a more holistic process that includes assessment of issues such as data quality, applicability of the model, and mechanistic interpretability in addition to statistical assessment.

The regulatory community is now attempting to provide a formal framework for the validation of QSARs [7]. Concepts within the framework can be related back to previous recommendations made within the field of QSAR. For instance, Tropsha et al. [8] described the requirements for QSAR models to be reliable and predictable. They men-

 DOI: 10.1002/qsar.200430909

tioned that models should be (1) statistically significant and robust (2) validated by making accurate predictions for external data sets that were not used in the model development and (3) have their application boundaries defined. They described also the statistical conditions for a QSAR model to be predictive, namely to have a coefficient of determination (r^2) > 0.6 and a (leave-one-out) cross-validated coefficient of determination (q^2) > 0.5.

More historically, in the early 1980s Unger and Hansch [9] stated that "*...without a quality perspective, one can generate statistical unicorns, beasts that exist on paper but not in reality*". To illustrate this point, Topliss and Costello [10] clearly showed that one could correlate a set of dependent variables using random numbers as independent variables. Later on, Kubinyi [11] emphasised the quality of fit and predictive ability of a regression model. Kubinyi [11] also summarised recommendations for the selection of appropriate regression model. The most important of these recommendations were proposed by Unger and Hanch in 1972 [9]. Among other criteria are (i) the principle of parsimony (Occam's razor) that the number of compounds per variable in the equation should be at least five to six to avoid chance correlation; (ii) the equation should be rejected if the number of variables in the regression equation is unreasonably large (i.e. the model is very complex); (iii) the standard deviation (standard error of estimate, s) should not be much greater than the mean error of the biological (toxicological) data.

Despite these formal recommendations there are few practical rules as to where to cease adding variables to a QSAR (i.e., what is an appropriate level of complexity). The complexity of any model is limited and therefore in reality will rarely match the complexity of the system being modelled. Fundamental research in this area confirms this point, for instance, Mikulecky [12] presented the emergence of complexity as an attribute of nature. He mentioned that there are 31 definitions of complexity, which were defined by Horgan [13]. Horgan [13] also pointed out the lack of a "unified theory" of complexity, simply because the entire real world is complex. The confusion between complication and complexity, and the difference between simulations and models require further elaboration. In the context of QSAR one should try to develop models which are simple (i.e., not complex) and for simple mechanisms.

Ideally, any QSAR, and especially those used for the regulatory purposes, should be derived using reliable and high quality data [14]. With regard to the complexity of QSARs, it is suggested that one should take the "simplest" model [11, 14]. The decision, however, is almost always subjective and there is little guidance for the modeller. The issue is complicated further when the multidisciplinary nature of the science is taken into account. The decision taken by the model developer might have different meaning for a statistician, chemist, toxicologist, or cheminformatician. Confronted with a large number of explanatory variables, the investigator interested in developing a regression model for the purpose of prediction is faced with two potentially conflicting objectives (1) selecting a model that will have the smallest prediction error and (2) estimating the predictive ability of the selected model. The desire for an economical choice of variables is motivated by the fact that the addition of a variable to a regression equation can, at times, increase the rates of prediction error. This latter phenomenon occurs as a result of the variability introduced by parameter estimation and is a function of the sample size. Unfortunately the information required to determine the best subset of variables is not immediately available to the investigator, although knowing which variables are relevant can give insight into the nature of the prediction problem.

A fundamental problem in the selection of the "best" model is that an appropriate balance between over-fitting and under-fitting of the model is required (i.e., the capability to scale up model complexity at the correct rate). Due to the inherent trade-off between model and data uncertainty, an optimal level of model complexity exists for every model. Methods for estimating data uncertainty are well known and routinely applied in most advanced risk assessment studies. This is seldom the case for model uncertainty. Models by their very nature are not identical to the reality they are meant to reflect. Sources of model uncertainty include assumptions, which have to be made to "reduce" reality to a model and the incomplete understanding of the physical phenomena underpinning the system behaviour. The theoretical relationship between model complexity and error is shown in Figure 1. This indicates that the error for the training set will decrease with increasing model complexity. However, the error for the test set (i.e., predictions from the model based on the training set) will reach a minimum with "moderate" model complexity and subsequently increase [15].

In the context of regression-based QSAR, it is well known that each additional descriptor included in a regres-

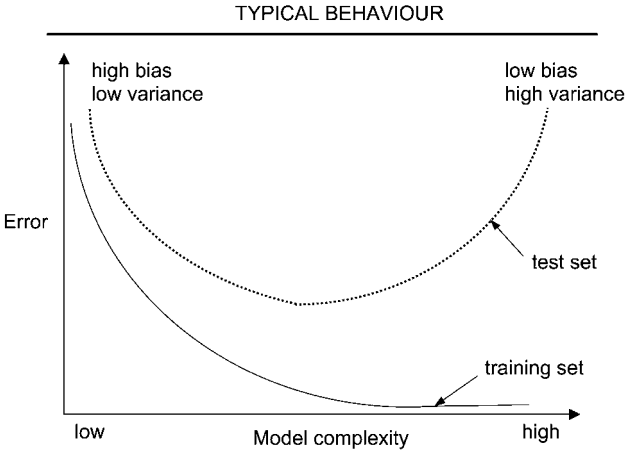

Figure 1. Theoretical profile of model behaviour

sion equation increases the variance explained. Thus, the fewer descriptors included, the less variance is explained. Unfortunately, using too few variables leads to an increased bias (i.e., the model is under-fitted). The emphasis in QSAR has traditionally been to achieve the greatest fit to the training set and hence the lowest standard error. However, in order to consider the potential of the model to predict toxicity, one should choose the simplest model that explains the data well and gives lowest error for test set (e.g., in terms of root mean square error of prediction, RMSEP). In other words, the emphasis should be shifted away from statistical fit and the prediction accuracy of model should be the "gold standard". The root mean square of error (RMSE) is calculated easily by squaring individual errors, summing them, dividing the sum by their total number, and then taking the square root of this quantity:

$$RMSE = \sqrt{\frac{\sum (Y^i_{pred} - Y^i_{\exp})^2}{N}} \quad (1)$$

The RMSE hence summarises the overall error of the model i.e. the precision of the QSAR and can thus be applied to predictions (i.e., RMSEP). It should also be remembered that every model contains simplifications such that predictions derived from the model will never be completely accurate and the model will never correspond exactly to reality. This concept is especially true when modelling as complex a phenomenon as biological activity.

The reliability and consistency of the experimental data in a modelling process are very important. The requirement for high quality data in QSAR is recognised as being paramount for successful modelling [14]. The TETRATOX database is considered to provide high quality acute toxicity data for modelling purposes [16]. Phenols form a large, structurally diverse and significant component of the TETRATOX database. They have been the subjects of a number of QSAR investigations [17, 18]. In addition, a number of mechanisms of action have been established for these compounds [19, 20]. Recently Cronin et al. [21] reported different QSARs for the prediction of the toxicity of enlarged data set of phenols to *Tetrahymena pyriformis*. The models reported by Cronin et al. [21] can be summarised as follow:

i) A model from response-surface (or two-parameter QSAR) analysis following the removal of outliers:

$$\log(IGC_{50})^{-1} = 0.53(0.022)\log D - 0.96(0.048)\,\mathrm{LUMO} - 0.58(0.057) \quad (2)$$

$n = 160$, $R^2 = 0.81$, $R^2_{CV} = 0.80$, $s = 0.34$, $F = 340$

ii) A model from stepwise regression analysis (no outliers removed):

$$\log(IGC_{50})^{-1} = 0.33(0.032)\log D - 0.45(0.078)\,\mathrm{LUMO} + 0.0028(0.0009)\,\mathrm{MW} - 0.020(0.0065)\,P_{NEG} + 0.036(0.011)\,\mathrm{SsOH} - 0.52(0.15)\mathrm{ABSQon} + 2.57(1.61)\,\mathrm{MaxHp} - 0.33(0.38) \quad (3)$$

$n = 200$, $R^2 = 0.65$, $R^2_{CV} = 0.63$, $s = 0.49$, $F = 54$

A model from stepwise regression analysis (outliers removed):

$$\log(IGC_{50})^{-1} = 0.38(0.024)\log D - 0.58(0.058)\,\mathrm{LUMO} + 0.0047(0.0008)\,\mathrm{MW} - 0.018(0.0048)\,P_{NEG} + 0.050(0.0083)\,\mathrm{SsOH} - 0.61(0.11)\mathrm{ABSQon} + 2.69(1.15)\,\mathrm{MaxHp} - 0.99(0.29) \quad (4)$$

$n = 185$, $R^2 = 0.83$, $R^2_{CV} = 0.82$, $s = 0.34$, $F = 128$

where D is the dissociation constant (i.e., the octanol-water partition coefficient corrected for ionisation), LUMO is energy of the lowest unoccupied molecular orbital, MW is molecular weight, P_{NEG} is the negatively charged molecular surface area, SsOH is the electrotopological state index for the hydroxy group, ABSQon is the sum of absolute charges on nitrogen and oxygen atoms and MaxHp is the largest positive charge on a hydrogen atom, n is the number of chemicals in the training set, R^2 is the coefficient of determination adjusted for the degrees of freedom and R^2_{CV} is the cross-validated (leave-one-out) coefficient of determination, s is the standard error of the estimate and F is the Fisher statistic.

The development of QSARs using different algorithms and computer software may fail to provide the globally optimal model, instead providing one that is locally optimal. Moreover, even a model with high correlation and cross-validated coefficients of determination may contain redundant descriptors or may not have a good predictivity. Therefore the two main goals of this work were firstly to describe some simple methods to select a precise model which, despite its uncertainties, can be used appropriately to make a decision; and secondly to determine if the RMSEP is a better fitness criterion than statistical fit in modelling.

2 Methods

2.1 Data set

The quality, consistency, and reliability of available data ultimately constrain the type and quality of model that can be applied. The TETRATOX database is a collection of toxic potency data to *Tetrahymena pyriformis* for more than 2,000 industrial organic compounds of which more than 1,500 have been published [16]. The present study utilises a subset of the TETRATOX database, which consists of toxicity data for approximately 250 phenols. The toxici-

ty of these chemicals was originally compiled and published by Cronin et al. [21]. The chemicals were divided into a training set of 200 chemicals and a test set of 50 chemicals. The validation set of phenols was selected prior to the development of QSAR models. The toxicity values themselves were measured in a population growth impairment test with the common ciliate *T. pyriformis* (strain GL-C), performed following the protocol described by Schultz [16]. The endpoint, population density, of this static 40-h assay was measured spectrophotometrically at 540 nm. Test conditions allow for 8–9 cell cycles in control cultures. Each compound was tested in a range finder prior to testing in duplicate for three definitive replicates. Two controls, one without test chemical but inoculated with *T. pyriformis*, and the other, a blank with neither test chemical nor ciliates, were used to provide a measure of the test acceptability and as a basis for interpretation of treatment data. Each definitive test replicate consisted of six to eight different concentrations with duplicate flasks of each concentration. Only replicates with control-absorbency values between 0.60 and 0.75 were used in the analyses.

The 50% growth inhibition concentration, IGC_{50}, was determined for each compound using the Probit Analysis routine in the Statistical Analysis System (SAS) software (SAS Institute 1989). The inverse of the logarithm of the millimolar concentration was used as the toxic potency endpoint.

Table 1. Chemicals tested, their toxicity to *Tetrahymena pyriformis* and significant physico-chemical descriptors.

ID	Name	CAS	MOA[a]	Toxicity	log *D*	E_{LUMO}	MW	P_{NEG}	ABSQon	MaxHp	SsOH
	Compounds for QSAR Development										
1	4-hydroxyphenylacetic acid	156-38-7	1	−1.50	−2.41	0.140	152.16	40.44	1.007	0.229	17.220
2	3-hydroxybenzyl alcohol	620-2-6	1	−1.04	0.30	0.388	124.15	47.97	0.784	0.210	17.380
3	4-carboxylphenol	99-96-7	1	−1.02	−1.36	−0.482	138.13	39.71	1.099	0.221	17.140
4	3-hydroxy-4-methoxybenzyl alcohol	4383-06-6	1	−0.99	0.00	0.338	154.18	42.21	1.121	0.219	17.860
5	4-hydroxy-3-methoxybenzylamine	7149-10-2	1	−0.97	0.01	0.266	153.20	38.43	1.053	0.181	9.160
6	4-hydroxyphenethyl alcohol	501-94-0	1	−0.83	0.62	0.333	138.18	35.58	0.754	0.217	17.390
7	3-carboxylphenol	99-06-9	1	−0.81	−1.77	−0.574	138.13	38.70	1.117	0.200	17.200
8	4-hydroxybenzamide	619-57-8	1	−0.78	0.23	−0.177	137.15	41.47	1.103	0.219	8.790
9	4-hydroxy-3-methoxybenzyl alcohol	498-00-0	1	−0.70	0.00	0.412	154.18	44.33	1.116	0.210	17.840
10	2,6-dimethoxyphenol	91-10-1	1	−0.60	0.77	0.388	154.18	42.64	1.105	0.177	9.340
11	2,4,6-tris(dimethylaminomethyl)phenol	90-72-2	1	−0.52	−0.75	0.425	265.45	32.71	1.438	0.217	10.380
12	Salicylic acid	69-72-7	1	−0.51	−2.28	−0.590	138.13	36.61	1.117	0.200	17.310
13	2-methoxyphenol	90-05-1	1	−0.51	1.19	0.391	124.15	41.60	0.754	0.173	8.990
14	5-methylresorcinol	504-15-4	1	−0.39	1.22	0.341	124.15	43.34	0.748	0.219	17.670
15	4-methylcyanophenol	14191-95-8	1	−0.38	0.71	0.063	133.16	45.98	0.566	0.217	8.840
16	3-hydroxyacetophenone	121-71-1	1	−0.38	1.38	−0.459	136.16	35.77	0.816	0.175	8.910
17	2-ethoxyphenol	94-71-3	1	−0.36	1.94	0.422	138.18	39.16	0.751	0.173	9.120
18	4-acetylphenol	99-93-4	1	−0.30	1.35	−0.380	136.16	38.63	0.793	0.220	8.830
19	3-ethoxy-4-methoxyphenol	65383-57-5	1	−0.30	1.78	0.318	168.21	41.16	1.097	0.179	9.120
20	2-methylphenol	95-48-7	1	−0.29	0.44	0.370	108.15	34.18	0.394	0.166	8.920
21	2-hydroxybenzamide	65-45-2	1	−0.24	1.37	−0.265	137.15	41.05	1.125	0.212	8.980
22	Phenol	108-95-2	1	−0.21	1.48	0.398	94.12	39.97	0.394	0.166	8.630
23	4-methylphenol	106-44-5	1	−0.18	1.94	0.431	108.15	36.58	0.394	0.166	8.760
24	4-hydroxy-3-methoxyphenethyl alcohol	2380-78-1	1	−0.18	0.33	0.324	168.21	36.72	1.188	0.209	17.860
25	3-acetamidophenol	621-42-1	1	−0.16	0.73	0.210	151.18	40.99	1.047	0.221	8.970
26	3-hydroxy-4-methoxybenzaldehyde	621-59-0	1	−0.14	0.98	−0.489	152.16	42.85	1.148	0.221	9.140
27	4-hydroxy-3-methoxyacetophenone	498-02-2	1	−0.12	1.32	−0.404	166.19	41.15	1.103	0.175	9.180
28	3,5-dimethoxyphenol	500-99-2	1	−0.09	1.42	0.415	154.18	46.71	1.094	0.220	9.100
29	2-hydroxyethylsalicylate	87-28-5	1	−0.08	1.52	−0.475	182.19	43.04	1.476	0.209	17.610
30	3-methylphenol	108-39-4	1	−0.06	1.94	0.394	108.15	37.11	0.394	0.166	8.810
31	Methyl-3-hydroxybenzoate	19438-10-9	1	−0.05	1.88	−0.485	152.16	45.04	1.105	0.177	8.950
32	3-methoxy-4-hydroxybenzaldehyde	121-33-5	1	−0.03	1.05	−0.478	152.16	41.73	1.148	0.221	9.090
33	4-hydroxy-3-methoxybenzonitrile	4421-08-3	1	−0.03	1.55	−0.429	149.16	48.16	0.954	0.172	9.100
34	3-ethoxy-4-hydroxybenzaldehyde	121-32-4	1	0.01	1.61	−0.452	166.19	39.95	1.157	0.181	9.220
35	4-fluorophenol	371-41-5	1	0.02	1.77	0.059	112.11	46.18	0.394	0.166	8.590
36	2-cyanophenol	611-20-1	1	0.03	1.21	−0.509	119.13	46.77	0.602	0.172	8.890
37	5-fluoro-2-hydroxyacetophenone	394-32-1	1	0.04	2.45	−0.786	154.15	40.83	0.738	0.173	9.020
38	2,4-dimethylphenol	105-67-9	1	0.07	2.40	0.399	122.18	37.69	0.394	0.166	9.040
39	2-hydroxyacetophenone	582-24-1	1	0.08	1.96	−0.517	136.16	36.51	0.748	0.168	9.060
40	2,5-dimethylphenol	95-87-4	1	0.08	2.40	0.347	122.18	35.50	0.394	0.166	9.100
41	Methyl-4-hydroxybenzoate	99-76-3	1	0.08	1.81	−0.397	152.16	41.62	1.086	0.221	8.860

Table 1. (cont.)

ID	Name	CAS	MOA[a]	Toxicity	log *D*	E_{LUMO}	MW	P_{NEG}	ABSQon	MaxHp	SsOH
42	3,5-dimethylphenol	108-68-9	1	0.11	2.40	0.387	122.18	32.21	0.394	0.166	8.990
43	4′-hydroxypropiophenone	70-70-2	1	0.12	1.91	−0.443	150.19	34.46	0.793	0.220	9.020
44	2,3-dimethylphenol	526-75-0	1	0.12	2.40	0.374	122.18	39.02	0.394	0.166	9.100
45	3,4-dimethylphenol	95-65-8	1	0.12	2.40	0.436	122.18	37.93	0.394	0.166	8.940
46	2-ethylphenol	90-00-6	1	0.16	2.47	0.386	122.18	35.87	0.394	0.166	9.110
47	Syringaldehyde	134-96-3	1	0.17	0.73	−0.505	182.19	44.45	1.454	0.179	9.440
48	Salicylhydrazide	936-02-7	1	0.18	0.58	−0.443	152.17	36.12	1.260	0.228	9.100
49	2-chlorophenol	95-57-8	1	0.18	2.01	0.030	128.56	39.39	0.393	0.166	8.790
50	4-hydroxy-2-methylacetophenone	875-59-2	1	0.19	1.83	−0.290	150.19	37.40	0.748	0.168	9.010
51	4-ethylphenol	123-07-9	1	0.20	2.47	0.435	122.18	30.65	0.394	0.166	8.850
52	3-ethylphenol	620-17-7	1	0.23	2.47	0.402	122.18	33.36	0.394	0.166	8.940
53	Salicylaldoxime	94-67-7	1	0.25	1.87	−0.312	137.15	44.58	0.903	0.208	17.140
54	2,3,6-trimethylphenol	2416-94-6	1	0.28	2.86	0.382	136.21	35.85	0.359	0.217	9.390
55	2,4,6-trimethylphenol	527-60-6	1	0.28	2.86	0.431	136.21	33.26	0.359	0.217	9.330
56	2-hydroxy-5-methylacetophenone	1450-72-2	1	0.31	2.42	−0.483	150.19	38.52	0.747	0.168	9.180
57	2-bromophenol	95-56-7	1	0.33	2.64	−0.049	173.01	36.02	0.394	0.166	8.870
58	5-bromo-2-hydroxybenzylalcohol	2316-64-5	1	0.34	1.31	−0.007	203.04	33.63	0.786	0.210	17.720
59	2,3,5-trimethylphenol	697-82-5	1	0.36	2.86	0.358	136.21	33.03	0.360	0.217	9.280
60	3-methoxysalicylaldehyde	148-53-8	1	0.38	1.34	−0.454	152.16	41.23	1.163	0.180	9.230
61	Salicylhydroxamic acid	89-73-6	1	0.38	0.47	−0.584	153.15	39.71	1.255	0.245	17.240
62	2-chloro-5-methylphenol	615-74-7	1	0.39	2.48	0.019	142.59	35.97	0.355	0.218	8.970
63	4-allyl-2-methoxyphenol	97-53-0	1	0.42	2.20	0.393	164.22	36.76	0.731	0.219	9.250
64	2-hydroxybenzaldehyde	90-02-8	1	0.42	1.55	−0.433	122.13	39.53	0.819	0.175	8.880
65	2,6-difluorophenol	28177-48-2	1	0.47	1.69	−0.321	130.10	45.75	0.379	0.175	8.460
66	Ethyl-3-hydroxybenzoate	7781-98-8	1	0.48	2.41	−0.453	166.19	42.38	0.973	0.181	9.020
67	4-cyanophenol	767-00-0	1	0.52	1.47	−0.413	119.13	47.37	0.602	0.172	8.740
68	4-propyloxyphenol	18979-50-5	1	0.52	2.37	0.330	152.21	36.45	0.732	0.219	18.270
69	4-chlorophenol	106-48-9	1	0.55	2.43	0.095	128.56	33.48	0.394	0.166	8.700
70	Ethyl-4-hydroxybenzoate	120-47-8	1	0.57	2.35	−0.367	166.19	39.73	1.083	0.221	8.920
71	5-methyl-2-nitrophenol	700-38-9	1	0.59	1.83	−1.153	153.15	31.03	0.359	0.217	9.100
72	2-bromo-4-methylphenol	6627-55-0	1	0.60	2.91	−0.012	187.04	32.46	0.392	0.167	9.000
73	2,4-difluorophenol	367-27-1	1	0.60	1.98	−0.318	130.10	40.10	0.379	0.176	8.500
74	3-isopropylphenol	618-45-1	1	0.61	2.82	0.415	136.21	31.28	0.394	0.166	9.060
75	5-bromovanillin	2973-76-4	1	0.62	1.39	−0.702	231.05	41.27	1.163	0.180	9.330
76	α,α,α-trifluoro-4-cresol	402-45-9	1	0.62	2.46	−0.348	162.12	39.49	0.394	0.166	8.660
77	Methyl-4-methoxysalicylate	5446-06-06	1	0.62	2.43	−0.428	182.19	45.92	1.424	0.179	9.340
78	4-bromophenol	106-41-2	1	0.68	2.49	0.020	173.01	34.76	0.394	0.166	8.740
79	2-chloro-4,5-dimethylphenol	1124-04-5	1	0.69	2.95	0.053	156.62	35.14	0.384	0.173	9.090
80	4-butoxyphenol	122-94-1	1	0.70	2.90	0.330	166.24	32.98	0.732	0.219	8.970
81	4-chloro-2-methylphenol	1570-64-5	1	0.70	2.89	0.080	142.59	30.52	0.394	0.166	8.990
82	3-*tert*-butylphenol	585-34-2	1	0.73	3.17	0.431	150.24	29.49	0.394	0.166	9.180
83	2,6-dichlorophenol	87-65-0	1	0.73	2.11	−0.259	163.00	31.36	0.388	0.169	8.940
84	2-methoxy-4-propenylphenol	97-54-1	1	0.75	3.00	−0.041	164.22	37.80	0.734	0.219	9.250
85	3-chloro-5-methoxyphenol	65262-96-6	1	0.76	2.64	0.027	158.59	38.27	0.749	0.175	8.960
86	4-chloro-3-methylphenol	35421-08-0	1	0.80	2.89	0.133	142.59	35.76	0.394	0.166	8.880
87	2-isopropylphenol	88-69-7	1	0.80	2.82	0.408	136.21	33.08	0.394	0.166	9.280
88	2,6-dichloro-4-fluorophenol	392-71-2	1	0.80	1.53	−0.568	180.99	25.83	0.380	0.175	8.900
89	4-iodophenol	540-38-5	1	0.85	2.91	0.024	220.01	34.63	0.394	0.166	8.750
90	2,2′-biphenol	1806-29-7	1	0.88	1.48	−0.239	186.22	42.72	0.788	0.166	8.630
91	4-*tert*-butylphenol	98-54-4	1	0.91	3.17	0.471	150.24	30.70	0.360	0.217	9.020
92	3,4,5-trimethylphenol	527-54-8	1	0.93	2.86	0.430	136.21	37.65	0.360	0.217	9.120
93	2,2′,4,4′-tetrahydroxybenzophenone	131-55-5	1	0.96	2.64	−0.786	246.23	45.66	1.923	0.221	8.850
94	4-*sec*-butylphenol	99-71-8	1	0.98	3.35	0.445	150.24	29.41	0.360	0.217	9.010
95	3-hydroxydiphenylamine	101-18-8	1	1.01	2.62	0.104	185.24	42.67	0.610	0.216	9.020
96	4-hydroxybenzophenone	1137-42-4	1	1.02	2.81	−0.485	198.23	40.89	0.744	0.167	9.100
97	2,4-dichlorophenol	120-83-2	1	1.04	2.91	−0.245	163.00	26.87	0.390	0.169	8.850
98	2,4,6-tribromoresorcinol	2437-49-2	1	1.06	2.74	−0.610	346.79	32.56	0.757	0.218	18.470
99	Benzyl-4-hydroxyphenyl ketone	2491-32-9	1	1.07	2.44	−0.375	212.26	40.24	0.750	0.166	18.280
100	4-chloro-3-ethylphenol	14143-32-9	1	1.08	3.42	0.141	156.62	29.57	0.390	0.170	9.010
101	2-phenylphenol	90-43-7	1	1.09	2.94	−0.119	170.22	40.93	0.359	0.217	9.560
102	2,5-dichlorophenol	583-78-8	1	1.13	2.66	−0.325	163.00	24.11	0.389	0.169	8.880

Table 1. (cont.)

ID	Name	CAS	MOA[a]	Toxicity	log *D*	E_{LUMO}	MW	P_{NEG}	ABSQon	MaxHp	SsOH
103	3-chloro-4-fluorophenol	2613-23-2	1	1.13	2.59	−0.264	146.55	38.47	0.383	0.174	8.690
104	3-bromophenol	591-20-8	1	1.15	2.62	−0.074	173.01	32.25	0.394	0.166	8.780
105	6-*tert*-butyl-2,4-dimethylphenol	1879-09-0	1	1.16	4.09	0.455	178.30	27.88	0.359	0.217	9.860
106	4-chloro-3,5-dimethylphenol	88-04-0	1	1.20	3.35	0.147	156.62	35.03	0.394	0.166	9.060
107	2-hydroxybenzophenone	117-99-7	1	1.23	3.39	−0.629	198.23	41.96	0.810	0.175	19.060
108	4-*tert*-pentylphenol	80-46-6	1	1.23	3.70	0.470	164.27	27.15	0.360	0.217	9.100
109	4-bromo-3,5-dimethylphenol	7463-51-6	1	1.27	3.41	0.109	201.07	32.55	0.395	0.166	9.100
110	4-bromo-6-chloro-2-cresol	7530-27-0	1	1.28	3.46	−0.226	221.48	33.75	0.393	0.167	9.180
111	4-cyclopentylphenol	1518-83-8	1	1.29	3.44	0.437	162.25	30.74	0.405	0.160	9.100
112	2-*tert*-butylphenol	88-18-6	1	1.29	3.17	0.436	150.24	31.61	0.394	0.166	9.450
113	2-*tert*-butyl-4-methylphenol	2409-55-4	1	1.30	3.63	0.477	164.27	31.02	0.394	0.166	9.570
114	2-hydroxydiphenylmethane	28994-41-4	1	1.31	3.47	0.242	184.25	38.66	0.360	0.217	9.310
115	Butyl-4-hydroxybenzoate	94-26-8	1	1.33	3.41	−0.367	194.25	35.89	1.083	0.221	9.370
116	3-phenylphenol	580-51-8	1	1.35	3.23	−0.161	170.22	40.68	0.360	0.217	9.270
117	n-pentyloxyphenol	18979-53-8	1	1.36	3.43	0.330	180.27	29.37	0.732	0.219	9.250
118	2,4-dibromophenol	615-58-7	1	1.40	3.31	−0.349	251.90	31.45	0.397	0.164	8.980
119	2,4,6-trichlorophenol	88-06-2	1	1.41	2.75	−0.502	197.44	21.69	0.385	0.171	9.010
120	2-hydroxy-4-methoxybenzophenone	131-57-7	1	1.42	3.43	−0.574	228.26	43.71	1.196	0.172	9.750
121	Isoamyl-4-hydroxybenzoate	6521-30-8	1	1.48	3.76	−0.363	208.28	34.01	1.083	0.221	9.570
122	3,5-dichlorosalicylaldehyde	90-60-8	1	1.55	2.41	−0.893	191.01	27.49	0.742	0.173	9.100
123	4-cyclohexylphenol	1131-60-8	1	1.56	4.00	0.442	176.28	29.22	0.360	0.217	9.140
124	3,5-dichlorophenol	591-35-5	1	1.57	3.25	−0.285	163.00	25.71	0.390	0.169	8.820
125	3,5-di-*tert*-butylphenol	1138-52-9	1	1.64	4.86	0.470	206.36	24.80	0.390	0.169	9.720
126	3,5-dibromosalicylaldehyde	90-59-5	1	1.64	2.67	−0.924	279.91	31.64	0.821	0.174	9.220
127	3,4-dichlorophenol	95-77-2	1	1.75	3.19	−0.236	163.00	29.66	0.390	0.169	8.790
128	4-bromo-2,6-dichlorophenol	3217-15-0	1	1.78	2.69	−0.514	241.89	25.89	0.389	0.169	9.050
129	2,6-di-*tert*-butyl-4-methylphenol	128-37-0	1	1.80	5.32	0.383	220.39	27.73	0.359	0.217	10.380
130	4-chloro-2-isopropyl-5-methylphenol	89-68-9	1	1.85	4.22	0.114	184.68	28.98	0.394	0.166	9.530
131	2,4,6-tribromophenol	118-79-6	1	2.03	3.28	−0.621	330.79	27.92	0.399	0.162	9.220
132	4-heptyloxyphenol	13037-86-0	1	2.03	4.50	0.329	208.33	27.82	0.732	0.219	9.070
133	4-*tert*-octylphenol	140-66-9	1	2.10	4.93	0.474	206.36	25.02	0.360	0.217	9.260
134	4-(4-bromophenyl)phenol	29558-77-8	1	2.31	3.95	−0.399	249.11	36.69	0.360	0.217	9.120
135	3,5-diiododsalicylaldehyde	2631-77-8	1	2.34	2.90	−0.901	373.91	34.31	0.708	0.194	9.280
136	2,3,5-trichlorophenol	933-78-8	1	2.37	2.84	−0.578	197.44	22.87	0.350	0.220	8.980
137	4-nonylphenol	104-40-5	1	2.47	6.19	0.429	220.39	24.65	0.360	0.217	9.150
138	Nonyl-4-hydroxybenzoate	38713-56-3	1	2.63	6.07	−0.368	264.40	29.71	1.083	0.221	9.120
139	2,4,6-trinitrophenol	29663-11-4	2	−0.16	−4.98	−2.534	229.12	30.49	0.359	0.217	9.130
140	3,4-dinitrophenol	577-71-9	2	0.27	0.24	−1.863	184.12	33.92	0.386	0.172	8.840
141	2,6-dinitrophenol	573-56-8	2	0.54	−1.67	−1.952	184.12	31.00	0.393	0.166	9.050
142	2,6-dichloro-4-nitrophenol	618-80-4	2	0.63	−0.66	−1.441	208.00	16.45	0.388	0.169	9.030
143	2,5-dinitrophenol	329-71-5	2	0.95	−0.16	−2.262	184.12	25.53	0.385	0.172	8.960
144	2,4-dinitrophenol	51-28-5	2	1.08	−1.57	−1.887	184.12	27.64	0.359	0.217	8.920
145	2,6-dinitro-4-cresol	609-93-8	2	1.23	−0.87	−1.893	198.15	31.03	0.359	0.217	9.170
146	4-bromo-2-fluoro-6-nitrophenol	320-76-3	2	1.62	0.13	−1.650	236.00	19.04	0.354	0.218	9.520
147	Pentafluorophenol	771-61-9	2	1.64	0.80	−1.296	184.07	1.90	0.370	0.181	8.300
148	4,6-dinitro-2-methylphenol	534-52-1	2	1.72	−0.73	−1.825	198.15	27.50	0.359	0.217	9.210
149	2,4-dichloro-6-nitrophenol	609-89-2	2	1.75	0.70	−1.579	208.00	15.72	0.384	0.172	9.060
150	Pentachlorophenol	87-86-5	2	2.05	2.11	−0.978	266.32	17.97	0.381	0.175	9.200
151	2,3,5,6-tetrachlorophenol	935-95-5	2	2.22	1.80	−0.817	231.88	21.35	0.383	0.173	9.140
152	Pentabromophenol	608-71-9	2	2.66	3.18	−1.193	488.57	27.07	0.403	0.160	9.520
153	2,3,4,5-tetrachlorophenol	4901-51-3	2	2.71	3.16	−0.752	231.88	21.36	0.384	0.173	9.050
154	4-acetamidophenol	103-90-2	3	−0.82	0.34	0.253	151.18	39.49	1.024	0.218	8.880
155	3-aminophenol	591-27-5	3	−0.52	0.34	0.522	109.14	46.32	0.684	0.162	8.730
156	4-aminophenol	123-30-8	3	−0.08	−0.29	0.439	109.14	42.89	0.684	0.162	8.700
157	3-methylcatechol	488-17-5	3	0.28	1.34	0.268	124.15	39.79	0.744	0.219	17.810
158	2-amino-4-*tert*-butylphenol	1199-46-8	3	0.37	2.13	0.418	165.26	29.75	0.677	0.164	9.180
159	4-methylcatechol	452-86-8	3	0.37	1.34	0.332	124.15	41.10	0.744	0.219	17.640
160	1,2,4-trihydroxybenzene	533-73-3	3	0.44	0.06	0.133	126.12	48.00	1.116	0.220	26.040
161	Hydroquinone	123-31-9	3	0.47	0.64	0.233	110.12	45.14	0.748	0.219	17.290
162	Catechol	120-80-9	3	0.75	0.88	0.297	110.12	42.31	0.744	0.219	17.340
163	2-amino-4-chlorophenol	95-85-2	3	0.78	1.67	0.043	143.58	34.89	0.681	0.162	8.860

Table 1. (cont.)

ID	Name	CAS	MOA[a]	Toxicity	log *D*	E_{LUMO}	MW	P_{NEG}	ABSQon	MaxHp	SsOH
164	1,2,3-trihydroxybenzene	87-66-1	3	0.85	0.28	0.029	126.12	45.53	1.113	0.220	26.090
165	2-aminophenol	95-55-6	3	0.94	2.44	0.406	109.14	38.19	0.681	0.162	8.790
166	4-chlorocatechol	2138-22-9	3	1.06	2.13	0.001	144.56	36.19	0.738	0.220	17.500
167	Chlorohydroquinone	615-67-8	3	1.26	1.51	−0.111	144.56	40.15	0.740	0.220	17.540
168	4-amino-2-cresol	2835-96-3	3	1.31	0.17	0.413	123.17	39.15	0.654	0.216	8.990
169	2,3-dimethylhydroquinone	608-43-5	3	1.41	1.56	0.215	138.18	40.41	0.747	0.219	18.230
170	4-amino-2,3-dimethylphenol	3096-69-3	3	1.44	0.63	0.406	137.20	39.25	0.681	0.163	9.170
171	Bromohydroquinone	583-69-7	3	1.68	2.00	−0.186	189.01	38.80	0.752	0.219	17.680
172	Tetrachlorocatechol	1198-55-6	3	1.70	3.07	−0.830	247.88	21.27	0.723	0.221	18.170
173	Phenylhydroquinone	1079-21-6	3	2.00	2.09	−0.229	186.22	44.77	0.750	0.219	18.860
174	3,5-di-*tert*-butylcatechol	1020-31-1	3	2.11	4.26	0.294	222.36	28.05	0.744	0.219	19.630
175	Methoxyhydroquinone	824-46-4	3	2.20	0.47	0.226	140.15	46.98	1.103	0.220	17.880
176	3-hydroxy-4-nitrobenzaldehyde	704-13-2	4	0.27	0.43	−1.755	167.13	26.22	0.751	0.168	8.990
177	5-hydroxy-2-nitrobenzaldehyde	42454-06-8	4	0.33	0.65	−1.486	167.13	29.79	0.787	0.186	8.870
178	2-amino-4-nitrophenol	99-57-0	4	0.47	0.59	−1.116	154.14	30.05	0.651	0.217	8.880
179	4-methyl-2-nitrophenol	119-33-5	4	0.57	1.92	−1.141	153.15	25.90	0.359	0.217	8.960
180	4-hydroxy-3-nitrobenzaldehyde	3011-34-5	4	0.61	−0.36	−1.456	167.13	26.28	0.751	0.168	8.940
181	4-nitrosophenol	104-91-6	4	0.65	0.51	−0.796	123.12	41.47	0.839	0.182	8.710
182	2-nitroresorcinol	601-89-8	4	0.66	−0.98	−1.321	155.12	21.13	0.747	0.219	17.720
183	4-methyl-3-nitrophenol	2042-14-0	4	0.74	2.37	−1.109	153.15	28.46	0.393	0.167	8.880
184	2-chloromethyl-4-nitrophenol	2973-19-5	4	0.75	0.73	−1.195	187.59	28.47	0.393	0.167	9.110
185	2-bromo-2′-hydroxy-5′-nitroacetanilide	3947-58-8	4	0.87	0.71	−1.105	275.07	22.41	1.041	0.232	9.210
186	4-amino-2-nitrophenol	119-34-6	4	0.88	0.53	−1.120	154.14	29.59	0.683	0.162	8.910
187	2-fluoro-4-nitrophenol	403-19-0	4	1.07	0.01	−1.333	157.11	24.13	0.353	0.219	9.010
188	5-fluoro-2-nitrophenol	446-36-6	4	1.13	0.76	−1.447	157.11	19.23	0.386	0.172	8.780
189	4-nitrocatechol	3316-09-04	4	1.17	1.05	−1.160	155.12	31.31	0.744	0.219	17.550
190	2-amino-4-chloro-5-nitrophenol	6358-07-02	4	1.17	2.38	−0.960	188.58	28.76	0.681	0.162	8.990
191	4-fluoro-2-nitrophenol	394-33-2	4	1.38	1.21	−1.447	157.11	25.27	0.354	0.219	8.930
192	4-nitrophenol	100-02-7	4	1.42	1.21	−1.065	139.12	26.69	0.394	0.166	8.720
193	2-chloro-4-nitrophenol	619-08-9	4	1.59	0.30	−1.264	173.56	23.75	0.393	0.166	8.870
194	4-chloro-6-nitro-3-cresol	7147-89-9	4	1.64	2.31	−1.346	187.59	25.47	0.394	0.166	9.090
195	3-methyl-4-nitrophenol	2581-34-2	4	1.73	1.74	−1.007	153.15	26.54	0.360	0.217	8.900
196	4-bromo-2-nitrophenol	7693-52-9	4	1.87	1.41	−1.398	218.01	30.38	0.361	0.217	9.040
197	4-chloro-2-nitrophenol	89-64-5	4	2.05	1.68	−1.388	173.56	18.73	0.394	0.166	8.910
198	Tetrabromocatechol	488-47-1	5	0.98	4.04	−0.987	425.68	29.28	0.759	0.217	18.630
199	Tetramethylhydroquinone	527-18-4	5	1.28	2.48	0.213	166.24	36.74	0.747	0.219	19.160
200	Tetrachlorohydroquinone	87-87-6	5	2.11	1.97	−0.928	247.88	23.22	0.725	0.221	18.300
	Compounds for QSAR Validation										
201	1,3,5-trihydroxybenzene	108-73-6	1	−1.26	0.05	0.247	126.12	50.41	1.120	0.220	26.022
202	2-hydroxybenzylalcohol	90-01-7	1	−0.95	0.30	0.344	124.15	37.22	0.783	0.210	17.509
203	Resorcinol	108-46-3	1	−0.65	0.76	0.321	110.12	44.46	0.748	0.219	17.306
204	4-(4-hydroxyphenyl)-2-butanone	5471-51-2	1	−0.50	0.93	0.319	164.22	33.74	0.756	0.168	17.379
205	3-methoxyphenol	150-19-6	1	−0.33	1.52	0.394	124.15	46.20	0.759	0.172	8.864
206	Ethyl-4-hydroxy-3-methoxyphenyl acetate	60563-13-5	1	−0.23	1.47	0.198	210.25	39.70	1.331	0.186	9.331
207	4-methoxyphenol	150-76-5	1	−0.14	1.31	0.303	124.15	46.44	0.759	0.172	8.797
208	3-cyanophenol	873-62-1	1	−0.06	1.68	−0.500	119.13	49.71	0.602	0.172	8.793
209	4-ethoxyphenol	622-62-8	1	0.01	1.84	0.327	138.18	40.58	0.756	0.172	8.869
210	4-hydroxypropiophenone	70-70-2	1	0.05	1.90	−0.364	150.19	34.80	0.793	0.220	8.900
211	3-hydroxybenzaldehyde	100-83-4	1	0.09	1.24	−0.547	122.13	39.23	0.820	0.175	8.785
212	4-chlororesorcinol	95-88-5	1	0.13	1.62	−0.008	144.56	38.09	0.740	0.220	17.527
213	2-fluorophenol	367-12-4	1	0.19	1.69	0.013	112.11	45.03	0.392	0.166	8.544
214	4-hydroxybenzaldehyde	123-08-0	1	0.27	1.24	−0.446	122.13	40.76	0.820	0.175	8.736
215	2-allylphenol	1745-81-9	1	0.33	2.50	0.348	134.19	39.70	0.394	0.166	9.194
216	3-fluorophenol	372-20-3	1	0.38	1.92	0.025	112.11	45.37	0.394	0.166	8.573
217	4-isopropylphenol	99-89-8	1	0.47	2.82	0.446	136.21	32.07	0.360	0.217	8.939
218	2-hydroxy-4-methoxyacetophenone	552-41-0	1	0.55	2.16	−0.455	166.19	42.47	1.105	0.175	9.292
219	3-methyl-2-nitrophenol	4920-77-8	1	0.61	1.66	−1.126	153.15	32.57	0.359	0.217	9.155
220	4-propylphenol	645-56-7	1	0.64	3.00	0.432	136.21	28.38	0.360	0.217	8.925
221	2-hydroxy-4,5-dimethylacetophenone	36436-65-4	1	0.71	2.88	−0.473	164.22	38.97	0.746	0.169	9.365
222	2-methyl-3-nitrophenol	5460-31-1	1	0.78	2.38	−1.090	153.15	28.64	0.360	0.217	8.967
223	3-chlorophenol	108-43-0	1	0.87	2.39	0.019	128.56	31.55	0.394	0.166	8.729

Table 1. (cont.)

ID	Name	CAS	MOA[a]	Toxicity	log *D*	E_{LUMO}	MW	P_{NEG}	ABSQon	MaxHp	SsOH
224	4,6-dichlororesorcinol	137-19-9	1	0.97	2.37	−0.263	179.00	29.48	0.734	0.220	17.748
225	4-benzyloxyphenol	103-16-2	1	1.04	2.96	0.232	200.25	40.93	0.728	0.166	9.086
226	3-iodophenol	626-02-8	1	1.12	2.92	−0.070	220.01	35.87	0.394	0.166	8.808
227	4-bromo-2,6-dimethylphenol	2374-05-2	1	1.17	3.41	0.085	201.07	34.29	0.361	0.217	9.310
228	2,3-dichlorophenol	576-24-9	1	1.28	2.61	−0.262	163.00	31.13	0.389	0.169	8.884
229	5-pentylresorcinol	500-66-3	1	1.31	3.35	0.345	180.27	30.06	0.748	0.219	18.381
230	4-phenylphenol	92-69-3	1	1.39	3.20	−0.086	170.22	41.48	0.360	0.217	9.104
231	Benzyl-4-hydroxybenzoate	94-18-8	1	1.55	2.77	−0.370	228.26	42.20	1.078	0.221	18.520
232	4-hexyloxyphenol	18979-55-0	1	1.64	3.97	0.330	194.30	28.51	0.732	0.219	9.043
233	4-hexylresorcinol	136-77-6	1	1.80	3.88	0.327	194.30	27.84	0.748	0.219	18.575
234	2,4,5-trichlorophenol	95-95-4	1	2.10	3.27	−0.555	197.44	25.58	0.386	0.171	8.950
235	2-ethylhexyl-4′-hydroxybenzoate	5153-25-3	1	2.51	5.34	−0.366	250.37	32.10	1.011	0.160	9.121
236	2,3-dinitrophenol	66-56-8	2	0.46	−0.03	−1.934	184.12	35.97	0.394	0.166	8.965
237	2,3,5,6-tetrafluorophenol	769-39-1	2	1.17	0.63	−0.994	166.08	13.28	0.367	0.183	8.338
238	2,6-diiodo-4-nitrophenol	305-85-1	2	1.71	−0.13	−1.422	390.90	30.53	0.373	0.180	9.275
239	3,4,5,6-tetrabromo-2-cresol	576-55-6	2	2.57	4.69	−0.882	423.71	30.72	0.402	0.161	9.565
240	2,4-diaminophenol	95-86-3	3	0.13	−1.80	0.527	124.16	40.75	0.971	0.161	8.863
241	5-amino-2-methoxyphenol	1687-53-2	3	0.45	−0.06	0.476	139.17	43.67	1.001	0.179	9.086
242	6-amino-2,4-dimethylphenol	41458-65-5	3	0.89	1.36	0.445	137.20	34.98	0.650	0.217	9.206
243	Trimethylhydroquinone	700-13-0	3	1.34	2.02	0.215	152.21	38.88	0.747	0.219	18.695
244	Methylhydroquinone	95-71-6	3	1.86	1.10	0.222	124.15	42.64	0.748	0.219	17.759
245	3-nitrophenol	554-84-7	4	0.51	1.89	−1.166	139.12	27.68	0.394	0.166	8.758
246	2-nitrophenol	88-75-5	4	0.67	1.29	−1.184	139.12	27.38	0.394	0.166	8.839
247	3-fluoro-4-nitrophenol	394-41-2	4	0.94	0.73	−1.284	157.11	24.85	0.355	0.219	8.657
248	2,6-dibromo-4-nitrophenol	99-28-5	4	1.36	−0.58	−1.453	296.90	24.33	0.396	0.164	9.196
249	4-nitro-3-(trifluoromethyl)-phenol	88-30-2	4	1.65	2.13	−1.585	207.12	33.09	0.360	0.217	8.760
250	Tetrafluorohydroquinone	771-63-1	5	1.84	1.38	−1.122	182.08	13.62	0.707	0.222	16.703

[a] Assigned mechanism of action (according to Aptula et al, 2002): 1 = polar narcotic; 2 = respiratory uncoupler; 3 = pro-electrophile; 4 = soft electrophile; 5 = pro-redox cycler

2.2 Calculation of Molecular Descriptors

The same 108 descriptors calculated previously from [21] were utilised in this study. Physico-chemical descriptors were calculated using a variety of software including the TSAR ver. 3.3 molecular spreadsheet (Accelrys Ltd., Oxford, England), the QSARis™ software ver. 1.1 (SciVision – Academic Press, San Diego, CA), the Chem-X ver. 2000.1 (Accelrys Ltd, Oxford, England) and ACD/Labs™ software (ACD/Labs™ software, **1995**, Advanced Chemistry Development Inc., Toronto, Canada). The distribution coefficient (log *D*) at pH = 7.35 was calculated according to the following expression:

$$\log D = \log P - \log\,(1 + 10^{\text{pH-pKa}}) \quad (5)$$

2.3 Model Development and Statistical Assessment

Models were developed using the same statistical software as applied by Cronin et al. [21] – the TSAR software package ver 3.3. (Accelrys Ltd., Oxford, England) and the MINITAB statistical software ver. 13.1 (MINITAB Inc., State College, PA). QSARs were developed on a data set of 200 phenols and validated using a test set of 50 chemicals. Various statistical methodologies were used for model development. Response-surface analyses were developed using MINITAB software and log *D* and LUMO as independent variables. The influence of number of chemicals in the training set on r^2 and q^2, the root mean square error of the calibration (RMSEC) and RMSEP were investigated. Stepwise regression analysis was performed to re-analyse the equations in order to investigate the influence of number of descriptors in MLR and q^2, RMSEC and RMSEP, respectively. Partial least squares analysis was performed using the TSAR software. Model fit was quantified with r^2 and q^2. The influence of the number of dimensions and the number of descriptors on RMSE were investigated.

3 Results and Discussion

3.1 Results from MLR

QSARs will be used increasingly to predict the toxicity and fate of chemical substances [22]. If QSARs are used for regulatory assessment, some estimation of the quality of the model is required, in particular in Europe [2]. Strategies for assessing the quality of models have been proposed [3, 7]. In addition, potential tools for the assessment of the quality of a model have been proposed [23]. More specifically, the individual quality of a model and its pre-

Table 2. Summary of physico-chemical descriptors calculated.

Software	Descriptors calculated
ACD/Labs	logarithm of the octanol-water partition coefficient ($\log P$), compound acidity (pK_a), molar refractivity (MR), molar parachor (PAR), molar polarisability (POL), surface tension (ST)
MOPAC (Chem-X)	energy of the highest occupied and lowest unoccupied molecular orbital (E_{HOMO}, E_{LUMO}), molecular electronegativity ($EN = -1/2\ (E_{HOMO} + E_{LUMO})$), hardness ($HARD = -1/2(E_{HOMO} - E_{LUMO})$), maximum acceptor ($Acc_{max}$) and donor ($Don_{max}$) superdelocalizability, maximum positive partial charge (Q_{max}), maximum positive partial charge on a hydrogen atom (Q_{Hmax}), maximum negative partial charge (Q_{min}), maximum diameter (D_{max})
Chem-X	volume, enclosed by isopotential surface with electrostatic potential (EP): $EP = -20$ kcal/mol (EP_{M20}); $EP = -10$ kcal/mol (EP_{M10}); $EP = 0$ kcal/mol (EP_{ZERO}); $EP = 10$ kcal/mol (EP_{P10}); $EP = 20$ kcal/mol (EP_{P20}). Coded by EP molecular VdW surface: $EP > 10$ kcal/mol (C_{POS}); $EP < -10$ kcal/mol (C_{NEG}); -10 kcal/mol $< EP <$ 10 kcal/mol (C_{MID}). Coded by EP molecular VdW surface, in percents: P_{POS}, P_{NEG}, P_{MID}
TSAR	dipole moment (μ), molecular volume (MV), molecular surface area (MSA), inertia moments (IM_1, IM_2, IM_3 size, and IM_1, IM_2 and IM_3 length), Wiener and Balaban topological indices, ellipsoidal volume (MV_{elipp}), molecular refraction (MR), number of hydrogen-bond donors (N_{Hdon}) and acceptors (N_{Hacc})
QSARis	Kier simple, valence and delta molecular connectivity indices ($^0\chi$, $^1\chi$, $^2\chi$, $^3\chi_p$, $^4\chi_p$, $^5\chi_p$, $^6\chi_p$, $^7\chi_p$, $^8\chi_p$, $^9\chi_p$, $^{10}\chi_p$, $^3\chi_c$, $^4\chi_{pc}$, $^6\chi_{pc}$, $^0\chi^v$, $^1\chi^v$, $^2\chi^v$, $^3\chi_p^v$, $^4\chi_p^v$, $^5\chi_p^v$, $^6\chi_p^v$, $^7\chi_p^v$, $^8\chi_p^v$, $^9\chi_p^v$, $^{10}\chi_p^v$, $^3\chi_c^v$, $^4\chi_{pc}^v$, $^6\chi_{pc}^v$, $\Delta^0\chi$, $\Delta^1\chi$, $\Delta^2\chi$, $\Delta^3\chi_p$, $\Delta^4\chi_p$, $\Delta^5\chi_p$, $\Delta^6\chi_p$, $\Delta^7\chi_p$, $\Delta^8\chi_p$, $\Delta^9\chi_p$, $\Delta^{10}\chi_p$, $\Delta^0\chi^v$, $\Delta^1\chi^v$, $\Delta^2\chi^v$, $\Delta^3\chi_p^v$, $\Delta^4\chi_p^v$, $\Delta^5\chi_p^v$, $\Delta^6\chi_p^v$, $\Delta^7\chi_p^v$, $\Delta^8\chi_p^v$, $\Delta^9\chi_p^v$, $\Delta^{10}\chi_p^v$), E-state indices ($SsCH_3$, $SssCH_2$, SdsCH, SaaCH, SaasC, $SsNH_2$, SddsN, SsOH, SdO, SssO, SsCl, CHHBd, SHBa), kappa and kappa alpha shape indices (^0k, ^1k, ^2k, ^3k, ^1ka, ^2ka, ^3ka), the sum of absolute values of the charge on each atom (ABSQ), the sum of absolute values of the charge on the nitrogen and oxygen atoms in a molecule ($ABSQ_{on}$), ovality, dipolar descriptors (Q_s, Q_v, Q_{sv})

dictivity are a basis for the decision of whether a model may be used for risk assessment. However, the natural complexity of toxicological and fate endpoints means that it is difficult to develop a mathematical model which will include all of the intrinsic mechanisms of relevant processes. Therefore, a model (e.g., QSAR for toxicity or fate) will always contain simplifications. As a result, predictions derived from these models can never be entirely accurate. Specifically with regard to QSARs, models validated internally (e.g., those that have been shown high r^2 and q^2), do not necessarily generate accurate predictions for new data [24].

To derive Eq. 3, a MLR equation for the toxicity of phenols to *T. pyriformis*, the authors [21] used 200 phenols in the training set. Seven descriptors were chosen by stepwise regression analysis: $\log D$, LUMO, MW, P_{NEG}, SsOH, ABSQon, MaxHp (as defined above). The resulting r^2 and q^2 were 0.65 and 0.63, respectively. To improve the statistical fit of this equation, fifteen of the greatest statistically significant outliers were deleted and 185 phenols were used to re-build the model. An r^2 and q^2 of 0.83 and 0.82, respectively, were subsequently obtained (Eq. 4).

In this study Eqs. 2 to 4 were re-analysed in order to investigate whether r^2 and q^2 are suitable criteria to determine the accuracy and the predictivity of QSAR models. In the current study several simple methods and criteria to develop better predictive models have been investigated. Firstly, the influence of the number of descriptors in model with regard to goodness of fit was analysed. Figure 2 presents the relationship between the number of descriptors in the MLR equations and q^2 (the number of chemicals in training set for first equation is 200 and 185 for the second). The results from MLR confirmed the well-recognised phenomenon that increasing the number of descriptors in a MLR increased r^2 and q^2 (see Figure 2).

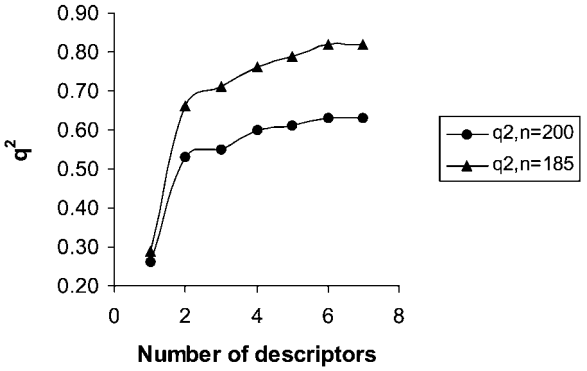

Figure 2. Plot of number of descriptors in MLR against q^2 for two data sets, $n = 200$ and $n = 185$

To investigate the influence of the number of additional descriptors added in MLR equations further, RMSEC and RMSEP were plotted against the number of descriptors (shown in Figure 3) for the two training sets ($n = 200$ and $n = 185$ chemicals) and test set ($n = 50$ chemicals).

Figure 3 shows that the RMSEC of the training set decreases with the increasing complexity of the model. In addition, the RMSEP of the test set appears to reach a local minimum after which it increases and as such could represent an overfitted model. This confirms the theory presented in Figure 1 and what can be expected when increasing the number of descriptors in a QSAR model. In this investigation the minimum RMSEP for the test set is for a

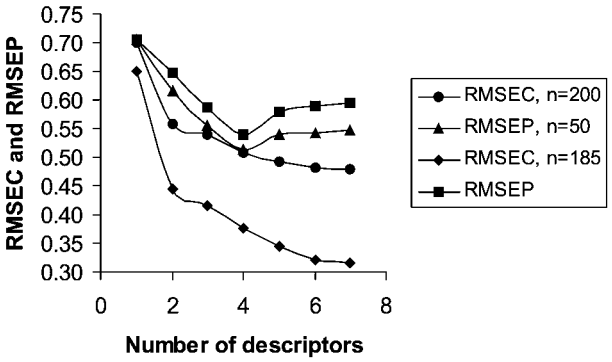

Figure 3. Plot of number of descriptors in MLR against RMSEC and RMSEP for two data sets, n = 200 and n = 185 and a test set, n = 50

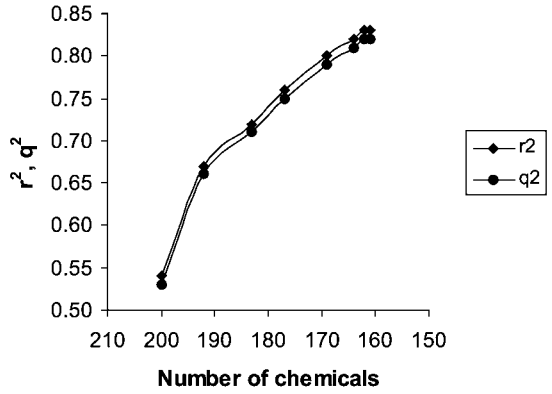

Figure 4. Plot of number of chemicals in two-parameter model against r^2 and q^2

MLR equation with four descriptors, namely log *D*, LUMO, MW and P_{NEG}. Interestingly, the RMSEP for the test set when the model is derived with 200 chemicals in training set is less (RMSEP = 0.514) than the RMSEP when there are fewer chemicals in the training i.e., 185 (RMSEP = 0.540). This clearly suggests that all available chemicals in the data set (i.e., training set) should be used for model development, regardless of whether they are statistical outliers.

A further modelling approach taken by Cronin et al. [21] was to develop QSARs using two descriptors, namely log *D* and LUMO. To improve the statistical fit of the equation derived using all 200 chemicals, the authors deleted 40 outliers (Eq. 2 above). Often in QSAR outliers are assumed to be erroneous data, or observations that cannot be explained [14]. However, outliers are commonly valid observations and as such they are the most interesting part of the data set. We re-investigated the influence of number of chemicals in the training set (i.e., the deletion of "outliers") on r^2 and q^2 for Eq. 2, the results are shown in Figure 4. As expected Figure 4 shows that r^2 and q^2 increase when "outliers" are deleted from the model. With

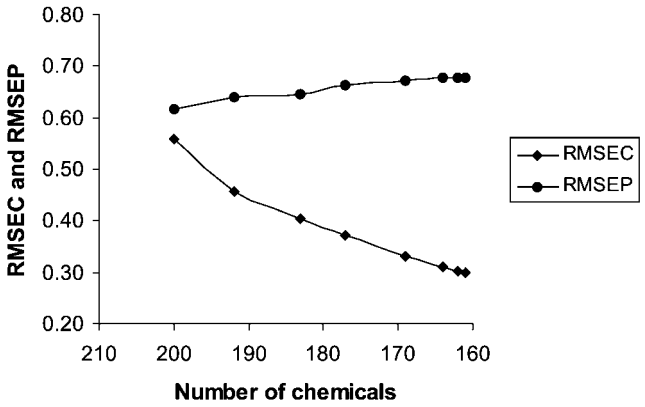

Figure 5. Plot of number of chemicals in two-parameter model against RMSEC and RMSEP

200 chemicals in the training set, r^2 and q^2 are 0.54 and 0.53, respectively. After the deletion of 40 chemicals (i.e., the outliers) from the training set, the statistics of the equation improve significantly with $r^2 = 0.81$ and $q^2 = 0.80$. The influence of the number of chemicals on RMSEC and RMSEP for the two-parameter QSAR is shown in Figure 5. The outliers were deleted from the training set after each step of the MLR modelling (a chemical was considered as an outlier if its standard residual > 0.2). Decreasing the number of chemicals following the deletion of outliers decreases the RMSEC for the training set which suggests a better fitting model, however, the RMSEP of the test set increases, being 0.617 for n = 200 and 0.679 when n = 160.

3.2 Results from PLS

Re-analysis of PLS model from [21] in which 14 descriptors were used (descriptors for PLS modelling were chosen after PCA), showed that only eight of them are required for the modelling of phenol toxicity. These eight descriptors were log *D*, LUMO, P_{NEG}, $^1\chi^v$, N_{Hal}, ABSQ, MaxNeg, SsOH, where $^1\chi^v$ is first order valence-corrected molecular connectivity index, N_{Hal} is number of halogen atoms in molecule and MaxNeg is the largest negative charge over the atoms in a molecule. PLS with these eight descriptors provided a two-dimensional model with r^2 0.76 and q^2 0.75. In the original work using all 14 descriptors r^2 and q^2 were 0.75 and 0.73, respectively. The influence of the number of dimensions in the eight descriptors PLS on the RMSE of the training and data set has also been analysed. The relationship between the number of dimensions in the PLS model and the RMSE of the training and test sets is shown in Figure 6. The results showed that r^2 and q^2 increase with increasing number of dimensions in the model (r^2 increased from 0.73 for one dimension to 0.79 for six dimensions; q^2 is increased from 0.71 to 0.77). As is expected the RMSEC for training set decreased, but the RMSEP for test set increased with the increasing number of dimensions (see Figure 6). This suggests that the RMSEP should

Table 3. Coefficients on individual variables and statistics of the eight parameter PLS model.

Descriptors	Dimension 2
log *D*	0.162
LUMO	−0.250
P_{NEG}	−0.027
$^1\chi^v$	0.316
N_{Hal}	0.158
ABSQ	−0.049
MaxNeg	4.752
SsOH	0.026
Constant	1.752
R^2	0.764
R^2_{CV}	0.745

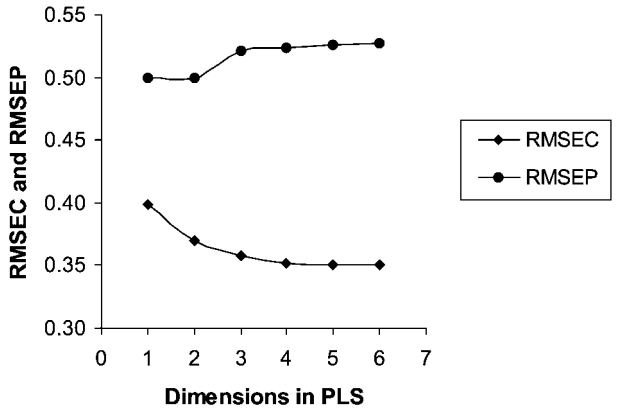

Figure 6. Plot of dimensions in PLS against RMSEC and RMSEP

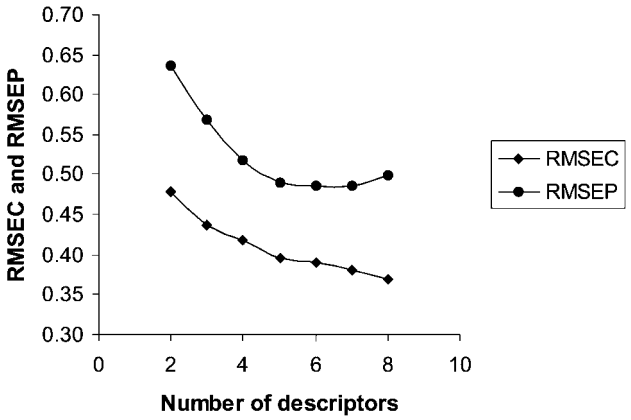

Figure 7. Plot of number of descriptors in PLS against RMSEC and RMSEP

be checked and the appropriate number of dimensions chosen for the PLS modelling.

To explore the effect of the complexity of the PLS model in terms of the number of descriptors and RMSEC and RMSEP, models with the eight descriptors noted above were used. The number of dimensions was set to two. The results shown in Figure 7 reveal the same tendency which was observed in MLR analyses. RMSEC for the training set decreases with the addition of the descriptors, while the RMSEP for the test set reaches a minimum value and then increases. The minimum value for RMSEP (0.485) was obtained when there were six descriptors in the PLS model.

4 Conclusions

A model may be developed for a variety of purposes. When it is used to predict ecotoxicity (or any toxicity or fate) the model should be valid, validated (in terms of q^2 and r^2 for test set as well as RMSEP) and easily interpretable. The results from this investigation showed that some important factors play a role in the definition of a good QSAR:

- q^2 is not a good criterion for a model predictivity.
- The number of compounds in the training set is important; do not delete "outliers" without well-defined reasons (e.g., a large leverage point). It should also be ascertained whether the data be more or less representative of the appropriate population if these data are deleted. If the outliers are true observations, systematically deleting them changes both the sample and the population of interest.
- The number of descriptors in the model (i.e., its complexity) is important for model under- and overestimation.
- An appropriate number of dimensions should be chosen for the PLS modelling.

When one develops or selects a model, it should be the model which will perform best on test data (i.e., the most predictive), as assessed by generalisation error i.e. RMSEP on these new data.

Acknowledgements

Dr Aptula gratefully acknowledges receipt of a Leverhulme Trust Fellowship.

References

[1] A. Combes, M. Balls, L. Bansil, M. Barrat, D. Bell, P. Botham, C. Broadhead, R. Clothier, E. George, J. Fentem, M. Jackson, I. Indans, G. Loizou, V. Navaratnam, V. Pentreath, B. Phillips, H. Stemplewski, J. Stewart, *ATLA* **2002**, *30*, 365–406.
[2] A. P. Worth, T. Hartung, C. J. van Leeuwen, *SAR QSAR Environ. Res.* **2004**, *15*, 345–358.
[3] A. P. Worth, C. J. van Leeuwen, T. Hartung, *SAR QSAR Environ. Res* **2004**, *15*, 331–343.

[4] M. T. D. Cronin, J. D. Walker, J. S. Jaworska, M. H. I. Comber, C. D. Watts, A. P. Worth, *Environ. Health Persp.* **2003**, *111*, 1376 – 1390.
[5] M. T. D. Cronin, J. S. Jaworska, J. D. Walker, M. H. I. Comber, C. D. Watts, A. P. Worth, *Environ. Health Perspect.* **2003**, *111*, 1391 – 1401.
[6] M. Balls, B. J. Blaauboer, J. H. Fentem, L. Bruner, R. D. Combes, B. Ekwall, R. J. Fielder, A. Guillouzo, R. W. Lewis, D. P. Lovell, C. A. Reinhardt, G. Repetto, D. Sladowski, H. Spielmann, F. Zucco, *ATLA* **1995**, *23*, 129 – 147.
[7] A. P. Worth, M. T. D. Cronin, C. J. van Leeuwen, A framework for promoting the acceptance and regulatory use of (Quantitative) Structure-Activity Relationships, in: M. T. D. Cronin, D. J. Livingstone (Eds.), *Predicting Chemical Toxicity and Fate*, CRC Press LLC, Boca Raton, Florida, **2004**, pp. 429 – 440.
[8] A. Tropsha, P. Gramatica, V. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69 – 77.
[9] S. Unger, C. Hansch, *J. Med. Chem.* **1973**, *16*, 745 – 749.
[10] J. Topliss, R. Costello, *J. Med Chem.* **1972**, *15*, 1066 – 1068.
[11] H. Kubinyi, *Quant. Struct.-Act. Relat.* **1994**, *13*, 285 – 294.
[12] D. Mikulecky, *Comput. Chem.* **2001**, *25*, 341 – 348.
[13] J. Horgan, *Scientific American*, **1995**, *June*, 104 – 109.
[14] M. T. D. Cronin, T. W. Schultz, *J. Mol. Struct. (Theochem)* **2003**, *622*, 39 – 51.
[15] T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer-Verlag, New York, **2001.**
[16] T. W. Schultz, *Toxicol. Methods* **1997**, *7*, 289 – 309.
[17] A. Aptula, T. I. Netzeva, I. V. Valkova, M. T. D. Cronin, T. W. Schultz, R. Kühne, G. Schüürmann, *Quant. Struct.-Act. Relat.* **2002**, *21*, 12 – 21.
[18] G. Schürmann, A. O. Aptula, R. Kühne, R.-U. Ebert, *Chem. Res. Toxicol.* **2003**, *16*, 974 – 987.
[19] T. W. Schultz, G. D. Sinks, M. T. D. Cronin, Identification of mechanisms of toxic action of phenols to *Tetrahymena pyriformis* from molecular descriptors. in: F. Chen, G. Schüürmann (Eds.), *Quantitative Structure-Activity relationships in Environmenatl Sciences – VII*. SETAC Press, Pensacola FL, **1997**, pp. 329 – 342.
[20] C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister, R. A. Drummond, *Environ. Toxicol. Chem.* **1997**, *16*, 948 – 967.
[21] M. T. D. Cronin, A. O. Aptula, J. C. Duffy, T. I. Netzeva, P. H. Rowe, I. V. Valkova, T. W. Schultz, *Chemosphere* **2002**, *49*, 1201 – 1221.
[22] M. T. D. Cronin, D. J. Livingstone (Eds.), *Predicting Chemical Toxicity and Fate*, CRC Press LLC, Boca Raton, Florida, **2004.**
[23] T. W. Schultz, T. I. Netzeva, M. T. D. Cronin, *SAR QSAR Environ. Res.* **2004**, 15, 385 – 397.
[24] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Model.* **2002**, *20*, 269 – 276.

