Molecular Diversity, **6:** 107–120.
KLUWER/ESCOM

Full Paper

Virtual generation of agents against *Mycobacterium tuberculosis*. A QSAR study

Emili Besalú[1]*, Robert Ponec[2] & Jesus Vicente de Julián-Ortiz[1]
[1] *Institute of Computational Chemistry, Universitat de Girona, Facultat de Ciències, Avda. Montilivi s/n, Girona, Spain;* [2] *Institute of Chemical Process Fundamentals, Czech Academy of Sciences, Prague 6, Suchdol 2, Czech Republic*
(Author for correspondence, E-mail: emili@iqc.udg.es, Fax: +34 972 418150)*

Received 28 April 2003; Accepted 3 June 2003

Key words: benzoxazine derivatives, cross-validation, Internal Test Sets method, linear models, phenylquinazoline derivatives, QSAR, statistical validation, tuberculosis, virtual molecular libraries

Summary

A QSAR approach based on the use of various topological indices as new theoretical molecular descriptors was applied to the study of a set of 64 anti-tuberculosis agents involving the substituted benzoxazines and phenylquinazolines. In order to evaluate the reliability of the proposed linear QSAR model, several statistical tests were proposed. The resulting model was subsequently applied to a wider virtual molecular library, which, together with the original set of 64 molecules with known activities contained another 512 molecules for which the predictions were made. Based on this prediction some new structures were proposed as especially promising candidates for active anti-tuberculotic drugs.

Abbreviations: ITS, Internal Test Sets; L1O, Leave-one-out; MIC, Minimal inhibitory concentration; MLR, Multilinear regression; QSAR, Quantitative Structure-Activity Relationships; TQSI, Topological Quantum Similarity Indices; UFS, Unsupervised Forward Selection.

Introduction

Tuberculosis, caused by *Mycobacterium tuberculosis*, kills more humans than malaria, AIDS, and all tropical diseases together. In recent years, the number of new cases worldwide has increased – one of the consequences of the AIDS epidemic. The dramatic increase in the number of new infections, which becomes the subject of important concern for public health, is due to two factors. The first is the resurgence of tuberculosis in the west from the 1980s, where the disease had been showing a steady decline from the beginning of the century [1]. The second, a number of outbreaks of multi drug resistant tuberculosis in many parts of the world, in the late 1980s and early 1990s. Infections caused by multidrug-resistant *M. tuberculosis* are difficult to treat. In addition, new *M. tuberculosis* strains have emerged that are resistant to all currently used anti-tuberculosis agents. These alarming trends prompted the World Health Organization to declare tuberculosis a global health emergency in 1993, an unprecedented distinction.

A review by World Health Organization of a series of 63 surveys of drug resistant tuberculosis carried out worldwide between 1985 and 1994 led to the conclusion that the new epidemic may be global [2]. Rates of primary resistance to isoniazid ranged from 0–17%; to streptomycin, 0–24%; to rifampicin, 0–3%; and to ethambutol, 0–4%. The rates of acquired resistance were isoniazid, 4-54%; streptomycin, 0–19%; rifampicin 0–15%; and ethambutol, 0–14%. Drug resistance in tuberculosis is not a new phenomenon. It was recognized very soon after the introduction of effective anti-tuberculosis drugs, that *M. tubercu-*

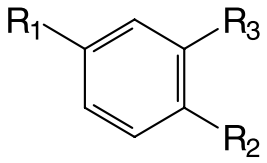


Figure 1. Substitution pattern defined in Combinator program in order to generate the virtual library.

losis could rapidly become resistant to the drugs used against it. Today, the term is used to signify disease due to *M. tuberculosis* that is resistant to the two most effective current anti-tuberculosis drugs, isoniazid and rifampicin, with or without resistance to other drugs [3].

Thus, there is an urgent need for potent inhibitors of *M. tuberculosis*, that exhibit favourable resistance profiles, and that are well tolerated by patients. New compounds potentially active against these bacteria are therefore constantly being sought [4, 5]. Most of the currently available drug design methods need a previous knowledge about the mechanism of action involved. However, extra-mechanistic approaches are increasingly used to the design of new drugs. Especially useful in this respect is the approach based on the exploitation the so-called topological indices as new molecular descriptors applicable for the design of theoretical Quantitative Structure-Activity Relationships (QSAR). The aim of this study is to develop new QSAR models able to predict *in vitro* activities of new drugs against *M. tuberculosis*. In this work, several benzoxazine and quinazoline derivatives with known anti-tuberculosis activity are studied in order to design new antibacterial compounds.

Molecular structures

Figure 1 depicts the general pattern used to identify all the molecules considered in this article. All of them can be formally seen as a substituted benzene ring at positions R_1, R_2 and R_3.

Waisser et al. have synthesised 64 structural derivatives belonging to 9 kinds of generic series and these investigators have also determined their respective activities against *M. tuberculosis*. Following the scheme of Figure 1, in Table 1 are found the original structures studied by Waisser. There, every generic label (**a, b, c, ...**) identifies a R_1 substitution. The activities [6–9] of these 64 compounds are shown in Table 2. *In vitro* activities are expressed as MIC (minimum inhibitory concentration in μmol L^{-1}) measured after 14 days of incubation, at 37 °C, of *M. tuberculosis* My 331/88 obtained from the Czech National Collection of Type Cultures, National Institute of Public Health, Prague. Reference compound is isoniazid, with MIC = 4 μmol L^{-1}. More details concerning the synthetic procedures and experimental protocols can be found in aforementioned studies. Molecular reference labels are obtained from Table 2. The label of a compound is the combination of a letter (from **a** to **i**) indicating the congeneric series in Table 1, and a number (from **1** to **12**) which will identify R_2 and R_3 substitutions. For instance, the parent non-substituted molecules of Table 1 (R_2=R_3=H) are denoted here as **a1, b1, ..., i1** and corresponding activities can be read in the first row of Table 2.

Virtual molecular generation and computation of indices

Based on the set of structures summarised in Table 2 a virtual library was generated in our laboratory. This was performed using Combinator program [10]. This computer code allows the generation of 3D chemical structures from a molecular basis or scaffold where different fragments are placed as substituents at different molecular sites. The general structure depicted in Figure 1 has been given to Combinator as the molecular generation basis.

In the next step the series of all the structures which can be formed by the combination of generation basis with all possible combinations of various R_1, R_2 and R_3 substituents was generated by the Combinator program. Taking into account that R_1 involves the set of 9 benzoxazine or phenylquinazoline moieties listed in Table 1, and each of the R_2 and R_3 corresponds to one of 8 substituents listed in Table 2, the above procedure resulted in the generation of the virtual library containing $9 \times 8 \times 8 = 576$ structures. This set involves as a subset the series of 64 molecules actually studied by Waisser, whose structures are also specified in the Table 2. After having generated the set of molecules, the geometries of all individual species were optimized using MOPAC [11] program.

Once the molecular library is created, each structure is sent to TQSI program [12], for the computation of indices. This program generates, for each molecule, several 2D topological indices (Wiener and Wiener Path Number, Randic, Schultz, Balaban and Hosoya indices, Harary Number, Kier and Hall generalised connectivity indices, and Gálvez charge indices)

Table 1. Generic compound series considered in this study. Each generic label identifies a R_1 substitution in Figure 1

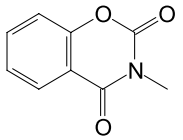
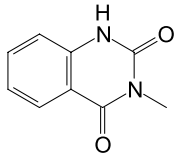
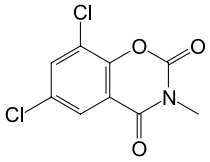
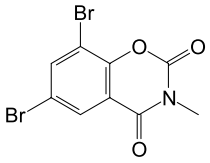
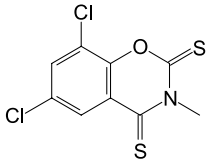
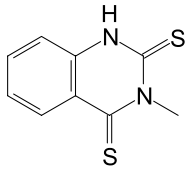
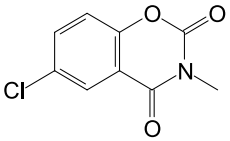
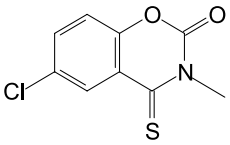
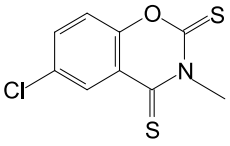
Generic label	Generic name	R_1 substitution in Fig. 1
a	3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-diones	
b	3-phenylquinazoline-2,4(1*H*,3*H*)-diones	
c	6,8- dichloro-3-phenyl-2*H*-1,3-benzoxazine-2,4(3H)-diones	
d	3-phenyl-6,8- dibromo-2H-1,3-benzoxazine-2,4(3*H*)-diones	
e	6,8-dichloro-3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-dithiones	
f	3-phenylquinazoline-2,4(1*H*,3*H*)-dithiones	
g	6-chloro-3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-diones	
h	6-chloro-3-phenyl-4-thioxo-2*H*-1,3-benzoxazine-2(3*H*)-ones	
i	6-chloro-3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-dithiones	

Table 2. Molecules tested by Waisser et al. Activities are expressed as MIC (in μmol L^{-1}). See text for more details

Number	Substituents		Congeneric series								
	R_3	R_2	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**
1	-H	-H	125	500	62	62	8	32	31	1	1
2	-H	$-CH_3$	62	500	16	31	4	16	16	0.5	0.5
3	$-CH_3$	$-CH_3$		250							
4	-H	$-OCH_3$	62	>125	125	62					
5	-H	-F	125	>125							
6	-Cl	-H	31		8	8	4		8	0.5	1
7	-H	-Cl	16		16	>16	8	32	4	0.5	0.5
8	-Cl	-Cl	8	>62	8	8		16	4	0.5	1
9	-H	-Br	31	250	16	31	16	16	4	0.5	0.5
10	$-NO_2$	-H	16	250		16					
11	-H	$-NO_2$	16								
12	-H	$-N(CH_3)_2$	>125		>62	>125					
No. of tested molecules			11	8	8	9	5	5	6	6	6
Reference			6	6	7	7	8	8	9	9	9

[13–16] and 3D variants constituting the so called Topological Quantum Similarity Indices (TQSI) [17, 18]. In this way, a set of 162 indices was obtained for each compound. More information about these indices, specially the quantum-related ones, can be found in the cited references.

The dimensionality of this primitive set of indices was reduced using the Unsupervised Forward Selection (UFS) algorithm by Whitley et al. [19]. UFS procedure has been implemented with a couple of modifications. Thus for example, while in the original algorithm, some vector descriptors were discarded if the absolute standard deviation from the mean is greater than a previously defined threshold, in our approach the descriptor variables were discarded if their *relative* variance was smaller than 5% compared to the mean value. In addition to this, the original procedure of discarding variables according to preselected parameter $R_{\max}$ was also modified so as to be less restrictive. As a consequence, the modified algorithm selects more indices than the original one of Whitley and collaborators. Thus, for example while original UFS procedure selected in our case the set of 11 descriptors, the modified procedure selected 66 different indices. Thus the final working matrix used in this work has dimension 576×66. In addition to this whole matrix a submatrix of dimension 64×66 was also considered in order to deal specifically with the 64 compounds of Table 2 for which experimental activities are available.

Tests of model predictability

A first test was performed on the series of 64 molecules of Table 2. It consisted in a standard leave-one-out (L1O) cross-validation procedure using Multilinear Regression (MLR) [20–22]. Models involving k descriptors (k = 1–6) were considered. For every value of k, the standard L1O procedure was performed using the Algorithm A, which also allows for the variable subset selection:

Algorithm A(*n, m, k*): Standard MLR-L1O on n molecules for obtaining linear models involving k indices selected from a set of m.

1. Generate all the $M = \binom{m}{k}$ combinations of k descriptors taken from the group of m.
   For every combination:
2. Perform a L1O test:
   - 2.1. For every one of the n molecules, left it apart and compute a MLR fitting equation involving the remaining $n - 1$ ones. Apply the obtained linear model to the excluded molecule, giving in this way a predicted value.
   - 2.2. Statistically evaluate the series of n predictions obtained in the previous step: com-

Table 3. Results obtained by the standard MLR-L1O procedure (Algorithm A) involving the 64 compounds with known activity. The confidence levels of all models are less than 0.0001

Number of descriptors (k)	Number of prediction series (M)	R_{cv}	$-\log P$
1	66	0.697	9.795
2	2145	0.852	17.17
3	45760	0.896	20.37
4	720720	0.918	22.33
5	8936928	0.936	24.29
6	90858768	0.944	24.90

Table 4. Results obtained for the 64 compounds with known activity and following the MLR-ITS-L1O method (Algorithm B). For each number of descriptors, and once all individual predictions are collected, the correlation coefficient between predicted and experimental quantities, R_{cv}, is tabulated. Pecka-Ponec statistical parameter is computed from this quantity

Number of descriptors	R_{cv}	$-\log P$
1	0.697	9.795
2	0.821	14.86
3	0.824	14.12
4	0.905	20.44
5	0.941	25.19
6	0.894	17.361

pute the correlation coefficient against the experimental values (R_{cv}) and the statistical significance of the attached single MLR fitted model involving the k descriptors (F-value, Student's t-values and significance for coefficients and independent term, etc.).

3. Final selected variables are those belonging to the combination having the highest R_{cv} coefficient and an acceptable statistical significance (level of significance <1%).
4. The final model then corresponds to the MLR fitting equation obtained considering all n molecules and the selected variables in the previous step.

In order to speed up the above reported L1O procedure in the step 2.1 of the algorithm A, the theorems [23] avoiding the explicit generation and solving of the whole set of n fittings involving $n - 1$ equations were used. Algorithm A was run for n = 64 molecules, m = 66 descriptors and $k = 1, 2, \ldots, 6$ descriptors using Regre, an in-house made program [24]. The results are summarised in the Table 3.

If models involving the same number of descriptors, k, are being compared, the best one will be the one having the highest R_{cv} correlation coefficient. In essence, this is the criterion followed in Algorithm A at step 3. Nevertheless, direct comparisons of this kind cannot be accomplished when considering models involving different number of descriptors, as those referred to in the Table 3. In order to overcome this problem, Pecka and Ponec [25] developed a fast and direct procedure for the comparison of statistical importance of the MLR correlation models differing in number of parameters and number of points. The method is based on the calculation of the probability that in a given correlation with n points and k parameters, the correlation coefficient higher than the one actually observed (R) can be obtained accidentally. This probability P is

$$P = \frac{\int_0^{\arccos R} \cos^{k-1}\theta \sin^{n-k-2}\theta d\theta}{\int_0^{\pi/2} \cos^{k-1}\theta \sin^{n-k-2}\theta d\theta}.$$

Intrinsically, Pecka-Ponec criterion is similar to a randomization test [26], because the evaluation of the integral takes into account the probability to obtain values of the correlation coefficient greater than the actual R. The lower the value of P, the more significant the model is. In this work, the negative logarithm of P is reported. As it is possible to see from the Table 3, the statistical importance of the QSAR models increases with increasing the number of descriptors (parameters) entering into the model. This, however, is not a general rule and in other situations a maximum in the value of $-\log P$ can often be found.

In connection with the linear cross-validation procedure implemented by means of Algorithm A, one has to be aware of the fact that in its original form this procedure usually over-estimates the predictive capabilities of selected models. The reason for that is simple: if a bigger pool of descriptors is considered, also the number of variable subsets, M, increases combinatorially, and the probability to find a best correlated series in Algorithm A at step 2.2 is higher. This, however, is just the situation, typical for unstable

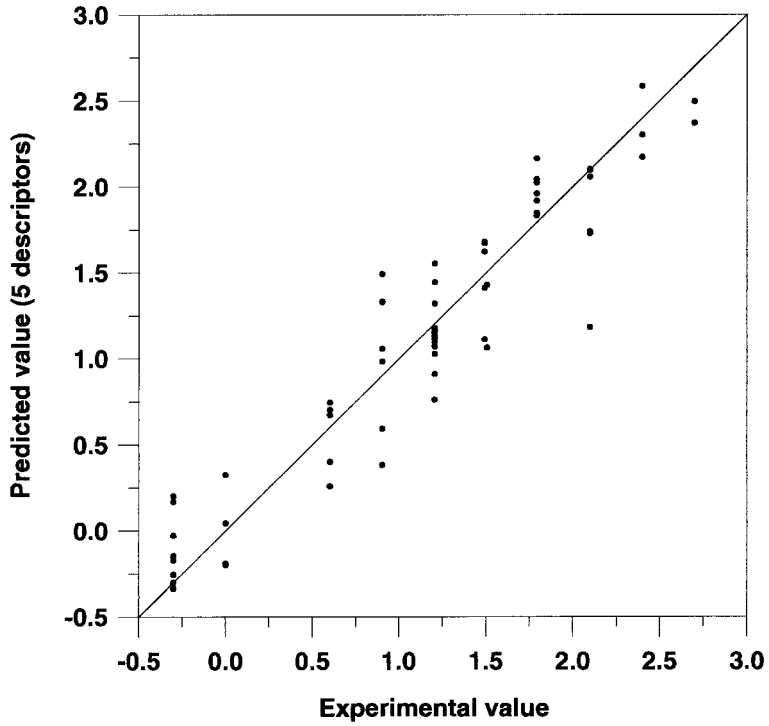


Figure 2. ITS-L1O results obtained for the 64 molecules with known activity and using 5 descriptors. Each point is attached to a fitted linear model.

over-parametrized models, which are difficult to reveal [27].

In order to overcome this drawback and to evaluate the real predictive capabilities of the selected models, an alternative cross-validation procedure called Internal Test Sets (ITS) Method [28] was considered by the authors. As it is explained in the literature, ITS method is a more realistic L1O procedure. In the case of linear models this method works as depicted in Algorithm B:

Algorithm B(*n, m, k*): MLR-ITS-L1O method for n molecules for obtaining linear models involving k indices taken from a set of m.

1. Consider the n molecules with known activity and left apart one at a time. For each molecule:

   1.1. Generate all the $M = \binom{m}{k}$ combinations of k descriptors taken from the group of m. For every combination:

      1.1.1. Obtain the MLR fitting equation involving the remaining $n - 1$ structures. Compute the correlation coefficient against the experimental values (R_{fit}) and the statistical significance indicators (F-value, Student's t-values and significance for coefficients and independent term, etc.)

   1.2. Final selected variables are those belonging to the combination having the best R_{fit} coefficient.

   1.3. Apply the MLR model which involves the variables selected in previous step to the molecule excluded in step 1 and obtain the property value prediction.

Algorithm B was run for the same set of molecules as in the previous case ($n = 64$, $m = 66$ and $k = 1, 2, \ldots, 6$ descriptors). An important feature of this alternative procedure is that for each particular value of k (denoting the number of parameters considered in the QSAR model) that Algorithm B always resulted in the generation of only one model for each molecule. In

Table 5. Indices involved in all the models selected by MLR-ITS-L1O procedure. The numerical coefficient sign was constant in all cases. See text for more details

Number of descriptors in model	Cardinal	Index	Times of usage in model	Percentage of intervention in equations	Coefficient sign
1	1	$^{9}\chi_{c}^{S}$	64	100	+
2	1	$^{4}J^{C}$	64	100	–
	2	$^{7}\chi_{c}^{C}$	35	54.7	+
	3	$^{9}\chi_{c}^{S}$	29	45.3	+
3	1	$^{7}\chi_{c}^{C}$	60	93.7	+
	2	$^{4}G^{C}$	58	90.6	–
	3	$^{6}G^{C}$	53	82.8	+
	4	$^{8}\chi_{ch}^{C}$	5	7.8	+
	5	$^{1}J^{S}$	3	4.7	–
	6	$^{5}J^{T}$	3	4.7	+
	7	$^{4}J^{T}$	3	4.7	–
	8	$^{1}G^{T}$	2	3.1	–
	9	$^{7}\chi_{c}^{T}$	2	3.1	+
	10	$^{9}\chi_{c}^{S}$	1	1.6	+
	11	$^{1}J^{C}$	1	1.6	–
	12	$^{9}\chi_{c}^{T}$	1	1.6	+
4	1	$^{4}G^{C}$	64	100	–
	2	$^{1}G^{S}$	55	86.0	–
	3	$^{8}\chi_{ch}^{S}$	55	86.0	+
	4	$^{7}\chi_{c}^{T}$	35	54.7	+
	5	$^{9}\chi_{c}^{T}$	20	31.2	+
	6	$^{1}J^{S}$	9	14	–
	7	$^{7}\chi_{c}^{C}$	9	14	+
	8	$^{8}\chi_{ch}^{T}$	9	14	+
5	1	$^{2}G^{S}$	64	100	–
	2	$^{7}G^{S}$	64	100	+
	3	$^{3}\chi_{c}^{C}$	64	100	–
	4	$^{7}\chi_{ch}^{C}$	64	100	+
	5	$^{9}\chi_{ch}^{C}$	64	100	+
6	1	$^{7}G^{S}$	63	98.4	+
	2	$^{2}\chi_{p}^{S}$	56	87.5	–
	3	$^{2}J^{S}$	44	68.8	–
	4	$^{6}\chi_{ch}^{T}$	37	57.8	–
	5	$^{7}\chi_{c}^{C}$	36	56.3	+
	6	$^{2}G^{S}$	20	31.3	–
	7	$^{9}\chi_{c}^{T}$	19	29.7	–
	8	$^{7}\chi_{c}^{T}$	15	23.4	–
	9	$^{5}G^{S}$	14	21.9	–
	10	$^{5}G^{T}$	13	20.3	+
	11	$^{8}\chi_{pc}^{S}$	13	20.3	+
	12	$^{7}\chi_{ch}^{C}$	12	18.8	+
	13	$^{3}\chi_{p}^{C}$	10	15.6	+
	14	$^{9}\chi_{ch}^{S}$	10	15.6	+
	15	$^{3}\chi_{c}^{C}$	3	4.7	–
	16	$^{9}\chi_{ch}^{C}$	3	4.7	+

Table 5. (continued)

Number of descriptors in model	Cardinal	Index	Times of usage in model	Percentage of intervention in equations	Coefficient sign
	17	$^{4}\chi_{p}^{T}$	3	4.7	–
	18	$^{8}\chi_{ch}^{S}$	2	3.1	+
	19	$^{4}\chi_{pc}^{S}$	1	1.5	+
	20	$^{7}J^{C}$	1	1.5	+
	21	$^{1}J^{T}$	1	1.5	–
	22	$^{7}\chi_{ch}^{T}$	1	1.5	+
	23	$^{5}\chi_{p}^{C}$	1	1.5	+
	24	$^{1}\chi_{p}^{S}$	1	1.5	–
	25	$^{3}J^{S}$	1	1.5	+
	26	$^{2}J^{T}$	1	1.5	+
	27	$^{9}\chi_{c}^{S}$	1	1.5	+
	28	$^{8}\chi_{ch}^{C}$	1	1.5	+
	29	$^{2}G^{T}$	1	1.5	+

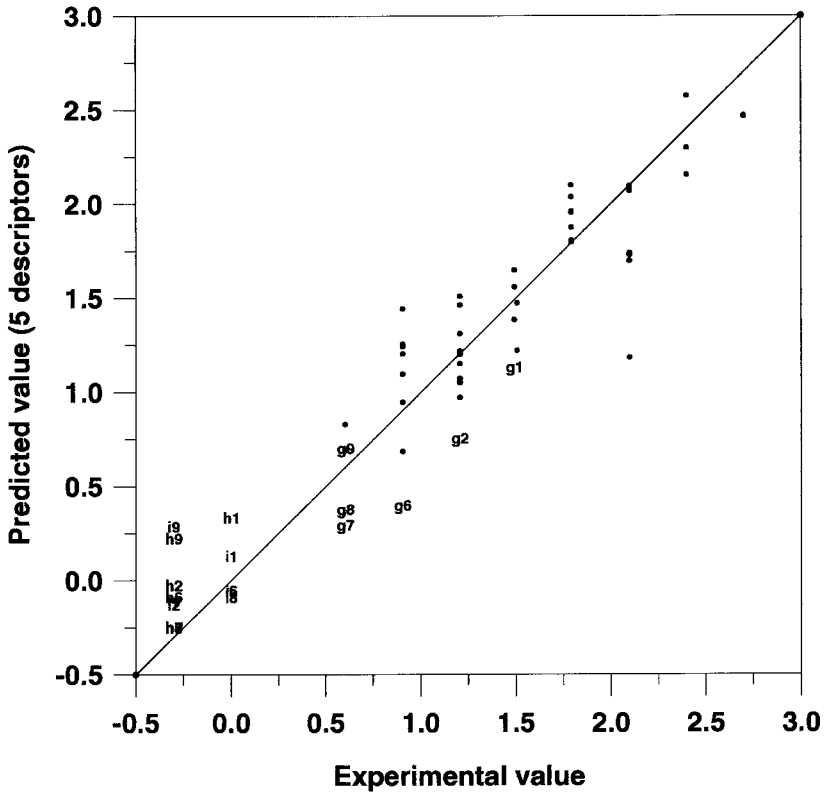


Figure 3. Molecule codes stand for predicted values against the experimental ones (see Table 6). Filled circles are fitted values in training process. See text for more details.

this way only a *single* prediction is made for the molecule that was left out. In other words, predictions are made at step 1.3 of Algorithm B without supervision. The main technical difference between the standard

Table 6. Predicted property values for the 18 molecules of groups **g, h** and **i**. Correlation coefficient respect experimental values is 0.875 and $-\log P = 2.757$, standard deviation is 0.306 logarithmic units

Number	Molecule	Actual logMIC	Predicted logMIC
1	**g1**	1.491	1.133
2	**g2**	1.204	0.756
3	**g6**	0.903	0.400
4	**g7**	0.602	0.294
5	**g8**	0.602	0.378
6	**g9**	0.602	0.700
7	**h1**	0.000	0.332
8	**h2**	–0.301	–0.026
9	**h6**	–0.301	–0.087
10	**h7**	–0.301	–0.246
11	**h8**	–0.301	–0.252
12	**h9**	–0.301	0.222
13	**i1**	0.000	0.127
14	**i2**	–0.301	–0.128
15	**i6**	0.000	–0.053
16	**i7**	–0.301	–0.121
17	**i8**	0.000	–0.090
18	**i9**	–0.301	0.285

L1O procedure (Algorithm A) and ITS-L1O one (Algorithm B) is evident. The loop procedures over the variable selection (generation of M subsets) and property prediction (keep a left out molecule) are *reversed*. This can be checked by looking at the respective algorithmic steps 1 and 2.1 from Algorithm A and 1 and 1.1 from Algorithm B. In this way, in an ITS-L1O procedure the process of selection of subset variables is performed without taking into account the information about the excluded structure. In practice, Algorithm B requires 64 independent calls of Regre program and, once a model is selected, the prediction for the hidden structure is obtained. This process of data hiding encompasses the recommendations of Hawkins [29] and collaborators who raise in the recent article the problem of real keeping the data in a cross-validation study, especially when the molecular set is small.

It is important to understand the difference between a standard linear L1O procedure and the ITS-L1O counterpart. In general, a reliable cross-validation procedure should deal with a successive series of training and true predictions. As a consequence, many QSAR studies deal with linear and non linear methodologies (principal component analysis, partial least squares, discriminant analysis, support vector machines, neural networks, classification techniques, etc., or combinations of them) implemented in such a way that the variable selection and the property prediction are performed in the same order as in Algorithm B. As a consequence the MLR-ITS-L1O linear procedure is the algorithm which should be preferred whenever the comparison of L1O cross-validated results with other methodologies is considered.

In both algorithms, the most inner loops involve the execution of MLR fittings. In order to see whether the linear models constructed at this level using leave-many-out procedures do indeed lead to improved results, the test calculations were performed. The result was negative and this is the reason why all the models presented in this paper come from ordinary MLR fitting procedures.

The statistical parameters obtained following the ITS-L1O method can be read in Table 4. As it is possible to see, the statistical importance of considered QSAR models (characterised by the value of –log P) increases with the increasing number of parameters only up to $k = 5$ but in contrast to the situation depicted in Table 3, further increase of the number of parameters leads to the decrease of the statistical importance of the corresponding 6-parameter QSAR models. So, the best models are the ones involving 5 descriptors. The predictions obtained from these models are depicted in Figure 2.

In connection with this conclusion it is, however, important to stress that straightforward comparison of the data in Tables 3 and 4 can be misleading. Thus, for, example, although some of the values of correlation coefficients obtained using original L1O procedure (Table 3) (not for $k = 5$) are higher than the ones from the Table 4, the conclusion that also the statistical importance of the results of the original L1O procedure is correspondingly higher would be wrong. This is due to the fact that tests of ITS-L1O procedure are much more severe than for standard L1O process and as each of 64 predictions presented in Figure 2 were obtained after a blind search, where the left out molecule data are completely hidden for the process, the true predictions are in fact performed only within ITS-L1O protocol.

In general, ITS procedures can be interpreted as being internal tests for assessing the true predictive capabilities of the model. Using an ITS method, detection of outliers is automatic [28], because only a single prediction is made by molecule on the fly and this prediction can differ considerably from the exper-

Table 7. Probabilities and significance levels to select from 0 to 10 or 12 active molecules (from a total of 12) when randomly selecting 10 or 20 compounds from the whole set of 64

Number of active molecules (p)	Selecting 10 molecules		Selecting 20 molecules	
	Probability $Q(p, 12; 10, 64)$ (%)	Significance level (%)	Probability $Q(p, 12; 20, 64)$ (%)	Significance level (%)
0	10.44	100.0	0.6422	100.0
1	29.15	89.56	4.670	99.36
2	32.79	60.41	14.35	94.69
3	19.43	27.62	24.61	80.33
4	6.653	8.189	26.15	55.72
5	1.359	1.536	18.09	29.58
6	0.1651	0.1771	8.331	11.49
7	0.0116	0.0120	2.563	3.157
8	0.0004	0.0004	0.5207	0.5940
9	<0.0001	<0.0001	0.0677	0.0733
10	*<0.0001*	*<0.0001*	0.0053	0.0056
11	–	–	0.0002	0.0002
12	–	–	*<0.0001*	*<0.0001*

imental value. On the other hand, in standard linear L1O or leave-many-out processes, outliers are difficult to detect due to possible over-fitting, resulting from the excessive number, M, of supervised prediction series.

There is another feature related to the conclusion that the best models are those involving 5 descriptors: a statistical study over the involved indices entering in all the models selected across the ITS-L1O involving from one up to six descriptors was performed. The frequency with which every index appears in models selected by Algorithm B are given in the Table 5. The interesting conclusion that can be deduced from the Table 5 is that the best models are the ones involving either one or five descriptors. These models are very stable because always the same indices have been selected in the step 1.2 of Algorithm B. In models involving 2 descriptors, the index ${}^{4}\mathrm{J}^{\mathrm{C}}$ is selected always while the pair of descriptors (${}^{7}\chi_{\mathrm{c}}^{\mathrm{C}}$ and ${}^{9}\chi_{\mathrm{c}}^{\mathrm{S}}$) appears with a slightly lower frequency. Similar inspection of the models involving 3 or 4 indices shows that there is again the set of indices, which enter into successful correlations very often. Moreover, the data in the table also suggests the lack of robustness of some linear models. This is especially true of the models involving 6 descriptors.

All in all, the above results allows us to conclude that the best linear models resulting from the reported search are those involving 5 descriptors. Moreover, the data in the Table 5 reveal that every time an index enters in a linear model, its numerical coefficient bears always the same sign. This is a very desirable property because this behaviour is suggesting for numerical consistency.

In order to provide another independent support for the above basic conclusion, an additional consistency test was performed. As we are interested in designing molecules having low value of logMIC, the following numerical experiment was designed. Let us choose as the working set the series of 46 test molecules (those of series **a, b, c, d, e** and **f**) and let the remaining 18 molecules belonging to groups **g, h** represent the validation set. The reason for the choice of this particular procedure is that 18 validation compounds taken from reference [9] showed a remarkable activity. MLR models involving from 1 up to 4 descriptors reasonably predicted the relative order of activities but all the calculated numerical values were greater than the actually observed ones (activity was underestimated). On the other hand, the model including 6 descriptors also gave a quite good sorting of the predicted activity, but all the numerical predictions were overestimated. The QSAR model which was able to predict the correct order of the activity of 18 test molecules and at the same time to reproduce negative values of logMIC was thus the above detected 5-parameter model. The actual form the 'best' 5-parameter QSAR model is given below.

Table 8. Ranking for the first molecules expected to have a notable activity against *M. tuberculosis*. R_1, R_2 and R_3 substitution places are those of Figure 1. Identification label refers to boxes in Table 2. See text for more details

Cardinal	–logMIC (expected)	Congeneric series (R_1)	R_2	R_3	Identification
1	–0.612	**h**	-$N(CH_3)_2$	-$NO(CH_3)_2$	
2	–0.536	**h**	-$N(CH_3)_2$	-NO_2	
3	–0.476	**h**	-Br	-Br	
4	–0.459	**h**	-$N(CH_3)_2$	-H	**h12**
5	–0.437	**i**	-Br	-Br	
6	–0.426	**h**	-NO_2	-H	**h11**
7	–0.397	**g**	-$N(CH_3)_2$	-$N(CH_3)_2$	
8	–0.393	**h**	-Br	-Cl	
9	–0.372	**i**	-NO_2	-H	**i11**
10	–0.371	**i**	-Br	-Cl	
11	–0.369	**h**	-Cl	-Br	
12	–0.332	**h**	-Cl	-Cl	**h8**
13	–0.328	**h**	-NO_2	-NO_2	
14	–0.317	**i**	-$N(CH_3)_2$	-H	**i12**
15	–0.300	**h**	-Br	-H	**h9**
16	–0.288	**i**	-Cl	-Br	
17	–0.263	**g**	-$N(CH_3)_2$	-NO_2	
18	–0.260	**i**	-Cl	-Cl	**i8**
19	–0.237	**i**	-H	-Br	
20	–0.185	**i**	-H	-Cl	**i6**
21	–0.184	**i**	-Br	-H	**i9**
22	–0.174	**i**	-Cl	-H	**i7**
23	–0.154	**h**	-Cl	-H	**h7**
24	–0.133	**g**	-NO_2	-NO_2	
25	–0.132	**h**	-Br	-CH_3	

$$\log\mathrm{MIC} = -(0.74352 \pm 0.31810)^{2}\mathrm{G}^{\mathrm{S}} + (3.1463 \pm 0.9225)^{7}\mathrm{G}^{\mathrm{S}} - (2.0010 \pm 0.3772)^{3}\chi_{\mathrm{c}}^{\mathrm{C}} + (4.3308 \pm 1.1686)^{7}\chi_{\mathrm{ch}}^{\mathrm{C}} + (4.5754 \pm 0.9638)^{9}\chi_{\mathrm{ch}}^{\mathrm{C}} - (14.962 \pm 3.825)$$

$$n = 46,\ R_{fitting} = 0.887,$$

$$-\log P = 11.73,\ F = 29.65\ .$$

Coefficient and independent term intervals are given at the 95% of confidence level.

Predicted activities presented in Table 6 summarise the results of the application of this model on the test set of 18 structures of series **g, h** and **i**.

The correlation of 18 predicted values (labelled points) with the corresponding experimental data are summarised in the Figure 3. For the sake of comparison, the same dependence also displays the data for 46 fitted values from the training set (filled circles).

In order to estimate the quality of the predictions obtained by the ITS methodology, another statistical significance test was designed. Results represented in Figure 2 were considered again and the set of 64 molecules was split into two subsets of molecules according to their activities. One of the subsets involved active and the other inactive molecules. A threshold value was set as follows: a molecule was labelled as active if the related MIC was lower than or equal to 1 μmol L^{-1} (logMIC$\leq$0). In this way the set of 12 active molecules, from the series **h** and **i** (see Table 2) was selected, the remaining molecules were considered as inactive.

The above sets of molecules were subjected to the following significance test. Let us imagine that we

have a series of s molecules from which we randomly select r molecules and we are asking what is the probability that p of the selected molecules out of q will be active. This probability is:

$$Q(p,q:r,s) = \frac{\binom{s-q}{r-p}\binom{q}{p}}{\binom{s}{r}} .$$

The results of this test are summarized in the Table 7, which lists the probabilities of having from 0 up to 10 active molecules (from the subset of 12) when choosing a random collection of 10 taken from the total of 64. The smaller is this value, the more effective is the selection. Right part of Table 7 shows the analogous data but calculated for the case of picking up form 0 up to 12 active compounds when selecting 20 molecules out of 64. Significance levels were obtained from cumulated probabilities: for a fixed value of p, they correspond to the probability to select p or more than p active molecules.

Let us confront now these results with the predictions summarised in Figure 2. Taking into account that the most active compounds are the ones with low MIC values and that the predicted activity is related to the vertical axis of the Figure 2, the structures with low predicted value on this axis can hopefully be regarded as the best candidates for preferential targets. In Table 7, cases marked in italics correspond to present results and, as it can be seen from the tabulated data, they are statistically very significant. This is because all 12 active compounds are found in the set of the first sorted 13 molecules.

Model for prediction

All the above reported tests clearly suggest that the best description of the studied series of molecules can be obtained by using a five parameter QSAR model. As a consequence and also according to Occam's razor philosophy, we confine ourselves in further considerations only to this particular model. The final proposed linear model arises from the correlation equation considering 5 variables selected by the Algorithm B and displayed in Table 5. This model was obtained by fitting the data of all the 64 compounds with known activity of Table 2:

$$\begin{aligned}\log\mathrm{MIC} = &-(0.75713 \pm 0.24380)^{2}\mathrm{G}^{\mathrm{S}} + \\ &(2.9196 \pm 0.6807)^{7}\mathrm{G}^{\mathrm{S}} - (2.0628 \pm 0.2520)^{3}\chi_{\mathrm{c}}^{\mathrm{C}} + \\ &(4.6406 \pm 0.7733)^{7}\chi_{\mathrm{ch}}^{\mathrm{C}} + (4.8861 \pm 0.5013)^{9}\chi_{\mathrm{ch}}^{\mathrm{C}} - \\ &(16.299 \pm 2.133)\end{aligned}$$

$$n = 64,\ R_{fitting} = 0.949,$$

$$-\log P = 27.10,\ F = 105.6 .$$

Coefficient and independent term intervals are given at the 95% of confidence level.

As expected, this model does not differ very much from the former one obtained by training 46 molecules of the groups **a, b, c, d, e** and **f**. Now, as all the 64 molecules enter into the fitting equation and the correlation coefficient increases, the statistical significance is also improved. This is properly reflected by the logarithmic Pecka-Ponec parameter (changing from 11.73 to 27.10). Note that, in this case, this statistical parameter is comparing models involving different number of points.

The proposed linear model was applied to all the 576 molecules from the virtual database originally generated by Combinator. This gave negative logMIC values for 46 molecules. Among them, there are 25 with logMIC ←–0.10 and 14 with logMIC <– 0.30. The first ranked molecules are listed in Table 8. All these molecules are proposed as potentially actives. All of them are variants of groups **g** and **h** (mainly) and **i**. This correlates with the data of Table 2. In fact, from the ranked sequence of predictions, the first molecule not belonging to one of these three groups is placed at position 53. It has to be taken into account that some ranked molecules (numbers 12,15,18,20–23) are training structures (**h8, h9, i8, i6, i9, i7, h7**, respectively), so the corresponding expected activities in Table 8 are fitted values. On the other hand, there are 4 compounds which represent real predictions because they correspond to empty boxes in Table 2, i.e. to molecules which were not yet prepared. These molecules specified as **h12, h11, i11**, and **i12** (ranking numbers 4, 6, 9, and 14), thus represent a real challenge for the synthesis and further testing as potential active targets. It is interesting that among the active structures proposed by the model, a special importance belongs to the molecules substituted by some groups from Table 2, namely -$N(CH_3)_2$, -NO_2, -Br, and -Cl. In some cases, di-substitutions are also proposed.

Conclusions

A QSAR study of a set of anti-tuberculosis agents has been performed. Special emphasis has been devoted to finely tuning the obtained linear model for prediction. Statistical tests for numerical stability and reliability of predictions have been pursuit. It has been shown that the reliability of the resulting model is crucially influenced by two factors. One of them is the stability of the model. This requires avoiding the eventual over-parametrisation or over-fitting of the model and one of the main advantages of the Algorithm B is that it gives account of this important factor. Another factor is the flexibility of the model, that should allow both, satisfactory description as well as reasonable extrapolation and predictions. This flexibility is included into the model via the limited number of parameters and the resulting 5-parameter model was found to satisfy reasonably both the above criteria. The overall predictive power was also evaluated by means of a statistical test (Q probability).

Acknowledgements

E. Besalú thanks the 'Generalitat de Catalunya' for an 'Agència de Gestió d'Ajuts Universitaris i de Recerca de la Generalitat de Catalunya' Grant No. 2002BEAI400110, which allowed visiting the Institute of Chemical Process Fundamentals in Prague, the place where this work was started. E. B. also thanks the Grant SAF2000-0223-C03-01 from the 'Ministerio de Ciencia y Tecnología. Plan Nacional I + D'. J. V. de Julián Ortiz acknowledges his Grant (EX2001-19851827) to the Spanish 'Ministerio de Educación'. Part of this work was performed during the visit of R. Ponec at the University of Girona supported by European Community project – Access to research Infrastructure action of the improving Human Potential Programme. This support is also gratefully acknowledged.

References

1. Kochi, A., *The global tuberculosis situation and the new control strategy of the World Health Organization*, Tubercle, 72 (1991) 1.
2. Cohn, D. L., Bustreo, F. and Raviglione, M. C. *Drug resistant tuberculosis: Review of the world-wide situation and the WHO/IUATLD Global Surveillance Project. International Union Against Tuberculosis and Lung Disease*, Clin. Infect. Dis., 24 (Suppl 1) (1997) S121.
3. Veen, J., *Drug resistant tuberculosis: Back to sanatoria, surgery and cod liver oil?*, Eur. Respir. J., 8 (1995) 1073.
4. Gestal-Otero, J. J., Figueiras-Guzmán, A. and Montes-Martínez, A., *Enfermedades infecciosas emergentes*. Med. Clin. (Barcelona), 109 (1997) 553–561.
5. Iseman, M. D. and Sbarbaro, J. A., *The increasing prevalence of resistance to antituberculosis chemotherapeutic agents: Implications for global tuberculosis control*, Curr. Clin. Top. Infect. Dis., 12 (1992) 188–204.
6. Waisser, K., Macháček, M., Dostál, H., Gregor, J., Kubicová, L., Klimešová, V., Kuneš, K., Palát, K., Hladůvkova, J., Kaustová, J. and Möllmann, U., *Relationships between the chemical structure of substances and their antimycobacterial activity against atypical strains. Part 18. 3-Phenyl-2H-1,3-benzoxazine-2,4 (3H)-diones and isosteric 3-phenylquinazoline-2,4(1H, 3H)-diones*, Collect. Czech. Chem. Commun., 64 (1999) 1902–1924.
7. Waisser, K., Hladůvková, J., Gregor, J., Rada, T., Kubicová, L., Klimešová, V. and Kaustová, J., *Relationships between the chemical structure of antimycobacterial substances and their activity against atypical strains. Part 14: 3-Aryl-6,8-dihalogeno-2H-1,3-benzoxazine-2,4(3H)-diones*, Arch. Pharm. Pharm. Med. Chem., 331(1) (1998) 3–6.
8. Waisser, K., Gregor, J., Dostál, H., Kuneš, K., Kubicová, L., Klimešová, V. and Kaustová, J., *Influence of the replacement of the oxo function with the thioxo group on the antimycobacterial activity of 3-aryl-6,8-dichloro-2H-1,3-benzoxazine-2,4(3H)-diones and 3-arylquinazoline-2,4(1H,3H)-diones*, Il Farmaco, 56(10) (2001) 803–807.
9. Waisser, K., Gregor, J., Kubicová, L., Klimešová, V., Kuneš, J., Machácek, M. and Kaustová, J., *New groups of antimycobacterial agents: 6-chloro-3-phenyl-4-thioxo-2H-1,3-benzoxazine-2(3H)-ones and 6-chloro-3-phenyl-2H-1,3-benzoxazine-2,4(3H)-dithiones*, Eur. J. Med. Chem., 35(7–8) (2000) 733–741.
10. Besalú, E., Combinator v1.3. Institute of Computational Chemistry, University of Girona, Spain, 2003.
11. Lobanov, V., MOPAC v6.00, University of Florida, 1996.
12. Besalú, E., TQSI program v1.0. Institute of Computational Chemistry, University of Girona, Spain, 2003.
13. A. T. Balaban, J. Chem. Inf. Comput. Sci., 35 (1995) 339–350.
14. Mihalic, Z. and Trinajstic, N., *A graph-theoretical approach to structure-property relationships (SYM)*, J. Chem. Educ., 69 (1992) 701–712.
15. Kier, L. B. and Hall, L. H., *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
16. Gálvez, J., Garcia, R., Salabert, M. T. and Soler, R., *Charge indices. New topological descriptors*, J. Chem. Inf. Comput. Sci. 34(3) (1994) 520–525.
17. Carbó-Dorca, R., Amat, L., Besalú, E., Lobato, M., 'Quantum Molecular Similarity', in R. Carbó-Dorca and P. G. Mezey (eds), *Advances in Molecular Similarity*, Vol. 2, JAI Press, London, 1998, pp. 1–42.
18. Besalú, E., Gironés, X., Amat, L. and Carbó-Dorca, R., *Molecular quantum similarity and the fundaments of QSAR*, Acc. of Chem. Res., 35 (2002) 289–295.
19. Whitley, D. C., Ford, M. G. and Livingstone, D., *Unsupervised forward selection: A method for eliminating redundant variables*, J. Chem. Inf. Comput. Sci., 40(5) (2000) 1160–1168.
20. Stone, M., *Cross-validatory choice and assessment of statistical predictions*, J. of the Roy. Stat. Soc. B, 36 (1974) 111–147.

21. Wold, S., *Validation of QSAR's*, Quant. Struct.-Act. Relat. 10 (1991) 191–193.
22. Wold, S. and Eriksson, L., 'Statistical Validation of QSAR Results. Validation Tools', in H. van de Waterbeemd (ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995, pp. 309–318.
23. Besalú, E., *Fast computation of cross-validated properties in full linear leave-many-out procedures*, J. Math. Chem., 29(3) (2001) 191–204.
24. Besalú, E., Regre v1.5. Institute of Computational Chemistry, University of Girona, Spain, 2003.
25. Pecka, J. and Ponec, R., *Simple analytical method for evaluation of statistical importance of correlations in QSAR studies*, J. Math. Chem., 27 (2000) 13–22.
26. Waller, C. L. and Bradley, M. P., *Development and validation of a novel variable selection technique with application to multidimensional quantitative structure-activity studies*, J. Chem. Inf. Comput. Sci., 39 (1999) 345–355.
27. Topliss, J. G. and Edwards, R. P., *Chance factors in studies of quantitative structure-activity relationships*, J. Med. Chem., 22(10) (1979) 1238–1244.
28. Besalú, E. and Vera, L., 'Internal Test Sets (ITS) Method: A New Cross-validation Technique to Assess the Predictive Capability of QSAR Models', in K. Sen (ed.), *Proceedings of the V Girona Seminar on Molecular Similarity, Girona (Spain)*, 12–20 July 2001, Nova Science Publishers, Inc., New York, 2003.
29. Hawkins, D. M., Basak, S. C. and Mills, D., *Assessing model fit by cross-validation*, J. Chem. Inf. Comput. Sci., 43(2) (2003) 579–586.