Legitimate Utilization of Large Descriptor Pools for QSPR/QSAR Models

Alan R. Katritzky,*[,§] Dimitar A. Dobchev,[‡,||] Svetoslav Slavov,[§] and Mati Karelson*[,‡]

Institute of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia, Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611, and Molcode, Ltd., Soola 8, Tartu 51014, Estonia

Received June 17, 2008

The use of large descriptor pools in multilinear QSAR/QSPR approaches has recently been increasingly criticized for their sensitivity to "chance correlations". Statistical experiments substituting "real descriptor" pools by random numbers were stated to demonstrate such sensitivity. While contributing positively to the improvement of the QSAR/QSPR methodology, these approaches claim complete interchangeability between the molecular descriptors used in QSAR/QSPR models and random numbers. Here, we demonstrate that when used correctly the large molecular descriptor pools are (i) not comparable with random numbers and (ii) can give very helpful QSPR conclusions.

1. INTRODUCTION

During the past decade, quantitative structure−activity relationships (QSAR) have been increasingly used in drug-design studies. Typically, a pool of independent variables (descriptors) of possible relevance to important physicochemical parameters relating to the series of compounds under discussion are evaluated by multiple-regression analysis for correlation with the activity values. The correlation equations which emerge from such analysis generally contain a small number of independent descriptors from the large pool evaluated. The descriptors selected for inclusion in such equations are chosen so that the overall relationships are highly significant by standard statistical criteria. However, these criteria relate to the individual variables in the final equation and do not take into account the number of descriptors actually screened for possible inclusion in the equation. Clearly, the larger the number of possible independent variables considered, the greater the possibility that a correlation will occur purely by chance. Because this factor is not reflected in the standard statistical criteria, it is important to consider the influence of the number of variables screened for possible inclusion in final equations.[1]

The utilization of QSAR/QSPR is now widely recognized as a proven and useful tool to elucidate the manner in which structure influences the behavior of properties or activities. Furthermore, it is now considered indispensable for the prediction of activities/properties of additional compounds. However, this widely ranging use is a relatively recent phenomenon, and proponents and the opponents of the method have both contributed significantly to the improvement of the methodology and reliability of the modeling procedures used. The complexity of QSAR/QSPR as an interdisciplinary science involving mathematics, chemistry, biology, and physics caused much discussion and controversy before the main concepts were formalized. A major controversial topic has been the possibility of "chance correlations" when using large descriptor pools.[2−4]

Recently, various sets of published multiparameter QSAR/QSPR models have been analyzed with particular attention turned to the possibility of "chance correlations" occurring in the published models.[5] We now demonstrate that "chance correlations" can be avoided by careful application of appropriate procedures for the selection of descriptors into a QSAR/QSPR model. We illustrate our treatment by using (i) a specific set of the antibacterial activities of 60 oxazolidinones (Set 10 in ref 5) and (ii) another data set comprising the vapor pressures of organic compounds.[6]

2. DATA SET

Rucker et al.[5] utilized a total of sixteen data sets to investigate "chance correlations" in QSAR/QSPR modeling and claimed that many of these sets had been treated by a stepwise algorithm with criteria for selection of the descriptors based only on R^2, a method that gave results no better than that obtained using random numbers. We have now demonstrated that this is not the case. In the present rebuttal, we consider general aspects related to the "chance correlations", using in particular the data set 10 from that reference.[5] This data set was used in our earlier work[7] to build three QSAR equations based on different descriptor spaces. We concentrate our current analysis on eq 1 of ref 7 which was developed using a large descriptor space comprising 1627 molecular descriptors. We then generalize the case and demonstrate that our conclusions given below are valid for the remaining eqs 2 and 3 in ref 7. We also removed one compound (#58) form the data set since it was a duplicate. Thus the original equation is slightly changed in terms of regression coefficients and statistical criteria. However, this correction does not influence the general conclusions drawn in the current work.

* Corresponding authors e-mail: katritzky@chem.ufl.edu (A.R.K.) and e-mail: mati.karelson@ttu.ee. (M.K.).
[§] University of Florida.
[‡] Tallinn University of Technology.
[||] Molcode, Ltd.

10.1021/ci8002073 CCC: $40.75
Published on Web 10/29/2008

3. METHODOLOGY AND DISCUSSION

3.1. Reliability of the Choice of Descriptor Scales Using the Stepwise Scale Selection Algorithm. In our original article,[7] the best multilinear regression (BLMR) algorithm[8] was used to build eq 1 by selecting m ($m = 7$) descriptors from the total pool of M (= 1627).

$$\text{Log}(1/\text{MIC}) = -164.88 + 151.63D_1 - 123.84D_2 + 9.76D_3 + 224.37D_4 - 18.16D_5 + 0.65D_6 + 0.03D_7$$

$$N = 60;\ n = 7;\ R^2 = 0.820;\ R^2_{cv} = 0.758;\ F = 33.77;\ s^2 = 0.082 \quad (1)$$

The seven descriptors involved in eq 1 are as follows: (i) average bond order of a H atom, D_1; (ii) minimum one-electron reactivity index for atom O, D_2; (iii) relative number of double bonds, D_3; (iv) minimum (>0.1) bond order for a N atom, D_4; (v) maximum sigma-π bond order, D_5; (vi) HOMO−LUMO energy gap, D_6; and (vii) total molecular one-center electron−electron repulsion, D_7. Each of these descriptors was discussed and related to the ligand-receptor interactions between the 3-aryloxazolidin-2-ones and the 50S ribosomal subunit of *S. aureus*.

The model of eq 1 was tested using both internal and external test set validation procedures. The QSAR model for the training set of 50 compounds was characterized by $R^2 = 0.809$. The validation set of 10 compounds produced $R^2 = 0.640$.

The BMLR method selects the best two-parameter regression equation, the best three-parameter regression equation, etc., based on the highest R^2 value in the stepwise regression procedure. During the standard BMLR procedure we use, the descriptor scales are normalized and centered automatically; the final result is given in natural scales. The following steps are carried out within this procedure to reduce the 1627 descriptors in the initial set to the 7 finally chosen:

(1) The following criteria were applied to eliminate all descriptors for which any of the conditions (a)-(f) applied: (a) The Fisher criterion (F) for the correlation between the descriptor and the property was above F_{min} (1). (b) The one-parameter equation correlation coefficient was less than our user-defined R_{min} value of $R_{min} < 0.1$ for insignificant correlations. (c) The level of significance (t) for the correlation between the descriptor and the property was less than $t = 1.5$. (d) The descriptor was intercorrelated ($R^2_{orth} \geq 0.2$) with another descriptor which possessed a higher single-parameter correlation coefficient value for the given property. (e) Descriptors that lead to nonsignificant 2-parameter equations ($R^2_{min} \leq 0.1$). (f) Descriptors which exhibited zero sample variance.

(2) The 400 two-parameter regression equations of orthogonal descriptor pairs possessing the highest R^2 value were submitted to the stepwise regression procedure described in (3) and (4) below.

(3) Each descriptor remaining in the pool (after the above deletions) that possesses coefficients R_{nc} below 0.8 for correlation with every descriptor already in the model was added, in turn, to the current n-parameter model. The resulting (n+1)-parameter models were tested, and the best 400 (n+1)-parameter models were resubmitted to the above procedure, to produce the best 400 (n+2) parameter models for a total of m iterations (m - defined by the user).

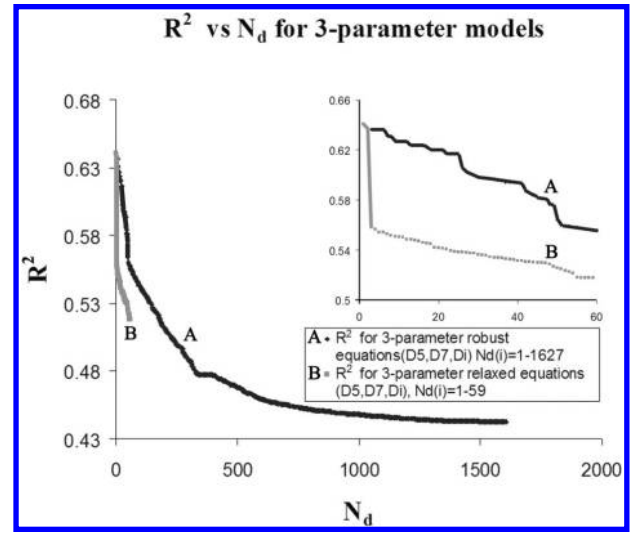

Figure 1. The R^2 values for 3-parameter QSAR models for the data on the antibacterial activity of oxazolidinones consisting of the combination of the descriptors in the best 2-parameter model with each of the remaining M-2 descriptors. N_d is the descriptor designation, ordered by the size of the R^2.

(4) The m parameter models with the highest F and R^2 values are finally analyzed.

(5) The "breaking point" rule based on the model plot of the R^2 vs number of descriptors is then used to select the optimal equation.

The article of Rucker et al.[5] claims that the QSAR models of similar, or better quality as estimated by the value of R^2, can be achieved by using random scales. We now re-examine this assertion using the current data set. We investigated the chance for a random descriptor occurring in eq 1 by estimating the possible number of descriptors out of the total number of 1627 descriptors that could enter into eq 1 when the above-described BMLR procedure and the adjacent criteria are applied.

The graph in Figure 1 gives the values of R^2 (coefficient of determination) of all 3-parameter models consisting of the combination of the descriptors in the best 2-parameter model ($R^2 = 0.443$, D_5, D_7) with each of the remaining M-2 descriptors (D_i). The graph in Figure 2 similarly gives the values of R^2 of all 5-parameter models consisting of the combination of the descriptors in the best 4-parameter model ($R^2 = 0.723$, D_3, D_4, D_5, D_7) with each of the remaining M-4 descriptors. The graph in Figure 3 gives the values of R^2 of all 7-parameter models consisting of the combination of the descriptors in the best 6-parameter model ($R^2 = 0.785$, D_1, D_3, D_4, D_5, D_6, D_7) with each of the remaining M-6 descriptors.

For a hierarchical comparison of the descriptors entering at each step into the 3-, 6-, and 7- models and thus in eq 1, we carried out i) *relaxed calculations* - calculations without using the criteria of our algorithm and ii) *robust calculations* - calculations applying all the criteria enumerated in steps 1−5 in the above-given description of the BMLR procedure. In each of Figures 1−3 N_d is the descriptor designation D_i ordered by R^2; the A lines in the figures represent relaxed, and the B lines robust calculations. It should be noted that for each of Figures 1−3, we performed two separate sets of calculations related to the (i) B (robust) and (ii) A (relaxed) lines. Each of the Figures 1−3 contains superimposed B and

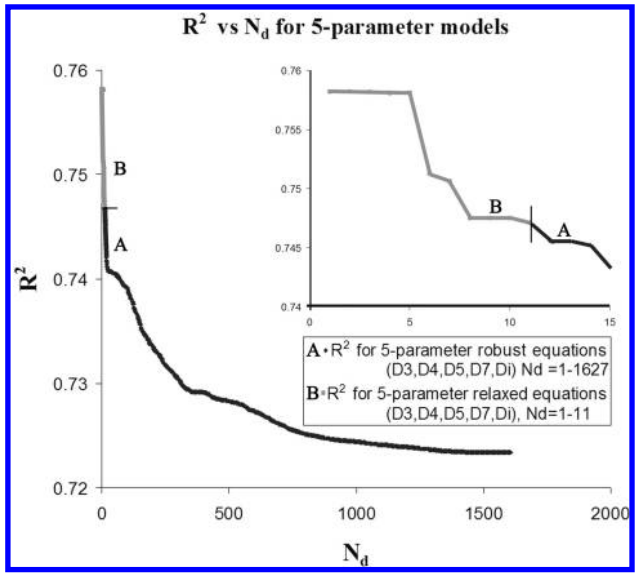

Figure 2. The R^2 values for 5-parameter QSAR models for the data on the antibacterial activity of oxazolidinones consisting of the combination of the descriptors in the best 4-parameter model with each of the remaining *M*-4 descriptors. N_d is the sequence order number of descriptors, ordered by the size of the R^2.

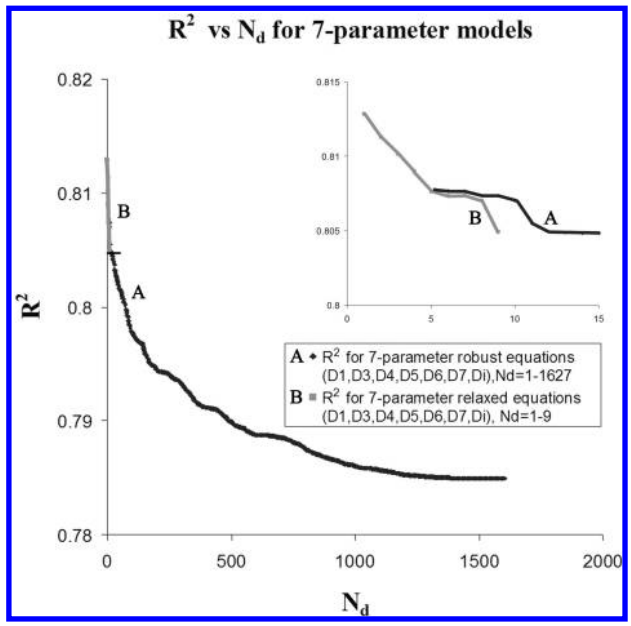

Figure 3. The R^2 values for 7-parameter QSAR models for the data on the antibacterial activity of oxazolidinones consisting of the combination of the descriptors in the best 6-parameter model with each of the remaining *M*-6 descriptors. N_d is the sequence order number of descriptors, ordered by the size of the R^2.

A lines which are not expected to coincide, because in a given R^2 interval there is a different number of descriptors (rejected or allowed by the criteria) ranked on the abscissa. Therefore, certain descriptors on a given abscissa position do not necessarily correspond to the same descriptors for the A and B lines. To show the differences between the B and A lines more precisely in terms of R^2, we presented additional "zoomed in" pictures attached to each Figure 1−3 (small figures). Each of the "small figures" demonstrates the gradual decrease of R^2 for n+1 correlations relatively to R^2 of the n correlations with the increase of n.

In the relaxed calculations (using no criteria, other than 1-f and the internal criteria for solving the normal equation), 1598, 1427, and 1587 descriptors enter the 3-, 5-, and 7-parameter models, respectively. In other words, 98%, 87%, and 97% of the total number of the descriptors take part in the BMLR procedure. As can be seen from the graphs, a significant number of descriptors give an improvement of the 2-, 4-, and 6- correlations whose number is defined as the total number of descriptors above R^2 0.443, 0.723, and 0.785, respectively (see Figures 1−3, A line). Therefore, the relaxed calculations indeed lead to results similar to the models built with random numbers in the article[5] in terms of selection criteria.

However, the results of the robust calculations, with the orthogonality constraints switched on, are quite different. Using the stepwise procedure with the criteria designated above for descriptor selection, the number of descriptors entering into the 3-, 5-, and 7- parameter models are now 59, 11, and 9 which corresponds to only 3.6%, 0.6%, and 0.55% of the total number of descriptors (1627). Thus for each multiparameter model, only a minor fraction of the more than 1600 descriptors available improve the correlation coefficient for the (n+1) parameter equation as compared with the corresponding n-parameter model. Most of the descriptors give no improvement and are thus statistically insignificant according to our criteria. Therefore, it is important to notice that the effective dimensionality of the descriptor space is much lower than the number of different descriptor scales in the initial pool. We note that the original eq 1 was built based on these "robust calculations" criteria.

3.2. The Orthogonality Problem. It is well-known that "chance correlations" can appear in using a multilinear regression procedure when the scales of independent variables are statistically closely related, i.e. *collinear*. Indeed, use of such scales for model development may produce statistically meaningless correlations with high R^2 values. The BMLR procedure specifically avoids these by elimination of the descriptors that are highly intercorrelated with the descriptors in the lower order model to which they are being added. The statistical experiment proving the above conjecture is given in the Supporting Information.

3.3. Randomization and Correlations with Random Numbers. In the critical article,[5] five different approaches (modes) to the randomization of the data related to multilinear QSAR regression models were applied using different combinations between the target variable (y, the property) and the independent variables (x, the descriptors). The combinations are noted as y vs x (p stands for permuted variable and r stands for replacement of the variable with random numbers): (i) mode 1 (*original target variable*) vs (*descriptors replaced by M random numbers*) - y vs rx; (ii) mode 2 (*original target variable randomly permuted*) vs (*M original descriptors*) - py vs x, also called y-randomization; (iii) mode 3 (*target variable replaced by random numbers*) vs (*M original descriptors*) - ry vs x; (iv) mode 4 (*target variable replaced by random numbers*) vs (*descriptors replaced by M random numbers*) - ry vs rx; and (v) mode 5 (*original target variable randomly permuted*) vs (*descriptors replaced by M random numbers*) - py vs rx. Because the authors of the above-mentioned work[5] did not have our original descriptors, they carried out calculations using only random numbers for the descriptors.

We have repeated the above experiments to show the results when using the original descriptors and the BMLR

Table 1. R^2 of the BMLR Equations for Mode 1, M=1627

equation	original	rand	genrand	statistica
2-descs	0.443	0.357	0.352	0.354
3-descs	0.641	0.478	0.478	0.482
4-descs	0.723	0.582	0.591	0.579
5-descs	0.745	0.675	0.682	0.679
6-descs	0.785	0.751	0.750	0.750
7-descs	0.813	0.806	0.804	0.799

Table 2. Averaged R^2 for All Order-Equations for Mode 2, M=7

equation	R^2
2-descs	0.142
3-descs	0.172
4-descs	0.188
5-descs	0.196
6-descs	0.204
7-descs	0.208

procedure described in section 3.1. For modes 1 through 5 we obtained the needed random numbers from 3 different sources: (a) standard C++ function; (b) genrand, [9] and (c) stat - STATISTICA 6.0 random number generator.[10] For reasons of diversity we used "rand" numbers generated in the range 0−32678 and for "genrand" and "stat" a range between 0 and 1. The generators were seeded with machine's system time at each run.

Mode 1. The mode 1 experiment used (a) 1627 and (b) 7 as the total numbers of descriptors M.

(a) For M = 1627, each trial was repeated 10 times, using the above-given BMLR criteria and procedure. We thus generated for each of 10 runs equations with up to 7 descriptors, from a descriptor-property matrix of dimensions 59 × 1628. Table 1 shows the results for the R^2 values from the three random number generators. Indeed, these results indicate significant correlations for the equations of all orders, especially for the higher order equations where the number of descriptor combinations for each equation is vastly larger than for the lower order equations. In this case, the results are close to those of the critical article's results[5] for the 7-descriptor equation using the corresponding models.

Nevertheless, comparison in Table 1 of the original equations with those simulated by using random numbers demonstrates that no linear combination of random descriptors correlates the real property better than the original descriptors. We note that this conclusion is valid only in this particular case and not universally true as it can depend on the dimensions of the descriptor-property matrix as well as the algorithm used to find the equations.

However, it is important to underline that all the randomly generated scales are practically orthogonal, and thus the space covered by them has much larger dimensionality than that of the original descriptors (many of which are nonorthogonal). We discuss this in more detail below as related to the treatment of a larger set of data on the vapor pressures of organic compounds.

b) Regarding the case M = 7, instead of the BMLR procedure, a direct stepwise MLR was applied in order to build all orders of the equations. Altogether 10000 7-descriptor sets were generated in this simulation. In no cases, at any order up to 7 was a significant equation found that exceeded $R^2 = 0.2$. Therefore, we conclude that the generation of large number of small sets consisting of 7 random variables cannot lead to multilinar models with $R^2 > 0.2$.

Mode 2. The mode 2 is one of the most important simulations whose results should be seriously taken into account when testing a real QSAR equation. The reason is that mode 2 shows the direct consequence when one intentionally breaks the inherited relationship between the property and the descriptors in the real equation. Moreover, if descriptors in a QSAR equation are physicochemically reasonable and the algorithm used to find the equation is sufficiently robust, then this QSAR relationship is expected to be unique for a given descriptor space. In other words, these descriptors do not appear by "chance". The terms "chance correlations" and "by pure chance" used by workers in the QSAR area are ambiguous, and their definitions are likely to depend on the knowledge of the researchers. In statistics, "chance" is always associated with the probability of a certain event to occur. The randomization tests applied in the present Article suggest a high probability of statistically insignificant descriptors in the equations relating i) final descriptors to the property and ii) the total descriptor space to the property. Cases (i) and (ii) are discussed below with M = 7 and M = 1627, respectively. At M = 7, the original descriptors for the 7-parameter equation reported in our work[7] were used, and at M=1627 all descriptors were involved.

Mode 2 at M = 7 is the standard randomization (y-randomization) to which special attention must be paid when one develops a reliable QSAR model. It directly assesses the significance of the final set of descriptors. Provided that one or several descriptors in the final equation are less statistically significant then the quality of the equations (in terms of R^2) would remain relatively high even after shuffling the y data. For example, in some cases where the final QSAR equation is overparametrized with a large number of descriptors, there is significant probability that some of the descriptors appear in the model because they (the descriptors) try to "fit" the random noise imposed by the experimental errors of the property. Note that we do not use the imprecise terms "chance correlations" or "chance descriptors" to describe some random events that do not depend on any measurable property. Instead we use the term probability for level of significance. The mode 2 experiment was repeated 1000 times as in each time (run) the order of the y values was randomly changed (scrambled) and the real descriptors were held unchanged. For each run the BMLR was used to generate models with up to 7-descriptors. The averaged R^2 was less than 0.21 for all orders of the models as can be noted from Table 2. Because of the similarity of the results, we present in Table 2 the averaged R^2 from the three random number generators. The resulting R^2 values (0.142−0.208) are low compared to the original models and especially the original 7-descriptor model ($R^2 = 0.81$). Therefore, according to this test, it is of low probability that the original 7-parameter model contains insignificant descriptors.

The case M = 1627 takes into account the full descriptor space. We note that this experiment does not directly assess the significance of the final equation's descriptors for a given QSAR model. The reason is that for each run the BMLR algorithm (or any other algorithm) finds different descriptors

Table 3. R^2 of the BMLR Equations for Mode 4, M=1627 [a]

equation	original	random
2-descs	0.443	0.371
3-descs	0.641	0.512
4-descs	0.723	0.604
5-descs	0.745	0.687
6-descs	0.785	0.753
7-descs	0.813	0.819

[a] The average R^2 is by 10, then by 3 for each generator.

Table 4. R^2 of the BMLR Equations for Mode 1, M=1627 for the Set of Data of Vapor Pressures of Organic Compounds

equation	original	random
2-descs	0.663	0.032
3-descs	0.816	0.046
4-descs	0.937	0.060
5-descs	0.940	0.073
6-descs		0.086
7-descs		0.098

for each run. This experiment is thus related to the problem of dimensionality of the data used to build the equations.

The mode 2 experiment with M = 1627 was repeated 100 times, and no significant QSPR equation was found. The BMLR procedure was able to find up to 2-descriptor equations with average $R^2 < 0.23$ only, and no further improvement was obtained by inclusion additional descriptor scales. This result was obtained when using the robust criteria described in section 3.1. It confirms that special care should be paid for the selection of the criteria of the algorithm used. For completeness, partial experiments were done where the criteria were relaxed to extreme cases (should not be done in practice) so as to force the BMLR to find higher order correlations. In no case did any of these 7-descriptor models provide R^2 exceeding 0.56−0.60; thus they were all inferior to the original.

Mode 3. Mode 3 is almost "equivalent" to Mode 2 at M = 7 and M =1627, and thus the results are quite similar, especially for the case of M = 7. In the case M = 7, the experiment differs from mode 2 by the range of the generated random numbers for the property by the intervals [0, 1] and [1, 32675]. There were slight differences in terms of R^2 when using only random numbers in the range [0, 1] - the maximum R^2 was 0.21 and for the range [1, 32675] - R^2 = 0.18.

The case with M = 1627 provided equations with only two descriptors; no higher order equations were found when using the robust criteria. The averaged R^2 from 10 runs for these 2 descriptor equations did not exceed 0.22.

Regarding the situation with the relaxed criteria, the few 7-descriptor models produced had $R^2 < 0.55$. This result indicates again that, in the absence of collinearity rejection criteria, use of a large number of descriptor scales has the potential to lead to "spurious" correlations.

Modes 4 and 5. Mode 4 is purely artificial, but it can illustrate trends related to the dimensions of the descriptor-property matrix used to build BMLR equations. The results depend on the quality of the random number generators. We assume the random number generators are of equal quality, and the conclusions depend only on the dimensions of the data used.

The results from the calculations of mode 4 are quite similar with those obtained in experiments with mode 1. At M = 1627, the number of trials was 10 for each generator, and the averaged R^2 values are given in the last column of Table 3. For comparison, the original R^2 of the real equations are also included in Table 3.

Again, it can be seen that the R^2 values are comparable with the R^2 values of the original QSPR models; these results were not sensitive to the criteria used for the BMLR. The random number vectors are mutually orthogonal; therefore, the criteria for orthogonality and collinearity rejected relatively few descriptors from inclusion in successive BMLR equations. Very similar results were obtained also for mode 5 at M = 1627. The same magnitude of R^2 resulted as in the last column of Table 3.

For both modes 4 and 5 at M = 7, the averaged R^2 for all orders of the equations produced were less than 0.19 and 0.183, respectively. Therefore conclusions for modes 4 and 5 at M = 7 are the same as for mode 1, case b.

Importantly the reliability of the multilinear QSPR models depends on both the size of the data set and the actual dimensionality of the descriptor space. Notably, many real descriptor scales are intercorrelated, and thus the actual space they cover has much lower dimensionality than the number of these scales.

Thus PCA analysis of the 1627 scales used in developing eq 1[7] produced just 19 principal components to explain 99% of data variance. The random scales are orthogonal by the definition. Therefore, it would have been correct to use only 19 random scales in the above-described numerical experiments with different modes. In such cases, of course, no significant correlations were obtained up to 7 parameter equations.

Moreover, the probability of chance correlations with random scales quickly diminishes with the increase of number of data points. In Table 4, we give the comparative R^2 values for successive best multilinear equations for the vapor pressures of organic compounds[6] obtained by using 645 data points and (i) the set of 946 natural descriptors originally calculated and (ii) the 1627 random scales, respectively. No significant model using random scales was detected by the BMLR procedure.

Two additional tests were carried out by us to demonstrate further the difference between real descriptor and random number scales.

First, we analyzed the statistical distribution of the random numbers (in the range 0..1) generated in Excel and the descriptors calculated by CODESSA PRO[11] (scaled in the same range). The results can be summarized as follows:

The means and the standard deviations of sets constructed from random numbers and "real" descriptors were calculated using STATISTICA. For sets of 10, 20, 30,..., 100 random numbers 5 trials were performed. The averaged mean values converged quickly to the expected mean value (when N tends to infinity) of 0.5 (see Table 5). The averaged standard deviation values similarly showed rapid convergence to near the expected value of 0.289. As seen from Table 5, when $N > 40$, the parameters characterizing the distribution of the random numbers are close to these for a uniform distribution.

However, the results were quite different (see Table 6) when a set of 100 randomly selected CODESSA PRO

Table 5. Means and Standard Deviations of Sets of Random Numbers

N	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean(av)	SD(av)
10	0.643	0.290	0.416	0.272	0.598	0.206	0.588	0.240	0.528	0.368	0.555	0.275
20	0.530	0.297	0.593	0.218	0.555	0.304	0.545	0.274	0.467	0.249	0.538	0.268
30	0.553	0.269	0.567	0.281	0.510	0.274	0.472	0.267	0.499	0.246	0.520	0.267
40	0.561	0.259	0.550	0.286	0.464	0.266	0.482	0.237	0.462	0.266	0.504	0.263
50	0.538	0.288	0.469	0.285	0.549	0.301	0.517	0.315	0.438	0.302	0.502	0.298
60	0.426	0.301	0.546	0.296	0.466	0.297	0.549	0.298	0.496	0.306	0.497	0.300
70	0.449	0.304	0.529	0.300	0.509	0.296	0.492	0.299	0.566	0.294	0.509	0.299
80	0.547	0.298	0.433	0.297	0.518	0.297	0.509	0.299	0.560	0.289	0.513	0.296
90	0.522	0.302	0.488	0.296	0.507	0.297	0.440	0.290	0.544	0.289	0.500	0.295
100	0.477	0.294	0.513	0.300	0.466	0.290	0.529	0.292	0.508	0.300	0.499	0.295

Table 6. Means and Standard Deviations of "Real" Molecular Descriptors

n	mean	SD	n	mean	SD	n	mean	SD	n	mean	SD	n	mean	SD
1	0.456	0.239	21	0.361	0.235	41	0.280	0.192	61	0.473	0.219	81	0.494	0.220
2	0.424	0.226	22	0.147	0.238	42	0.451	0.218	62	0.460	0.222	82	0.541	0.232
3	0.495	0.238	23	0.230	0.423	43	0.453	0.234	63	0.237	0.165	83	0.410	0.251
4	0.117	0.249	24	0.877	0.175	44	0.413	0.229	64	0.401	0.192	84	0.512	0.249
5	0.140	0.377	25	0.141	0.244	45	0.364	0.197	65	0.591	0.181	85	0.040	0.024
6	0.020	0.141	26	0.091	0.172	46	0.431	0.211	66	0.532	0.200	86	0.194	0.156
7	0.080	0.419	27	0.190	0.273	47	0.602	0.233	67	0.339	0.187	87	0.084	0.144
8	0.410	0.866	28	0.125	0.240	48	0.438	0.187	68	0.304	0.298	88	0.072	0.139
9	0.047	0.157	29	0.175	0.253	49	0.473	0.221	69	0.322	0.197	89	0.047	0.033
10	0.020	0.141	30	0.146	0.278	50	0.464	0.225	70	0.555	0.176	90	0.474	0.305
11	0.334	0.088	31	0.294	0.154	51	0.267	0.176	71	0.397	0.191	91	0.442	0.281
12	0.553	0.138	32	0.184	0.220	52	0.443	0.209	72	0.283	0.162	92	0.271	0.327
13	0.160	0.249	33	0.274	0.232	53	0.634	0.195	73	0.071	0.117	93	0.386	0.232
14	0.010	0.028	34	0.497	0.229	54	0.378	0.226	74	0.087	0.137	94	0.449	0.245
15	0.018	0.124	35	0.474	0.243	55	0.318	0.219	75	0.081	0.126	95	0.479	0.215
16	0.013	0.076	36	0.518	0.217	56	0.259	0.192	76	0.496	0.238	96	0.540	0.202
17	0.085	0.195	37	0.352	0.255	57	0.604	0.204	77	0.451	0.212	97	0.334	0.196
18	0.048	0.169	38	0.445	0.186	58	0.490	0.221	78	0.448	0.193	98	0.537	0.167
19	0.002	0.012	39	0.479	0.203	59	0.326	0.187	79	0.486	0.223	99	0.346	0.255
20	0.434	0.224	40	0.439	0.197	60	0.434	0.186	80	0.406	0.170	100	0.741	0.214

molecular descriptors were studied. For compatibility in the statistical calculations, all descriptor values were scaled in the same range (0..1) used for the random numbers. Even at $N = 100$, only 14 out of 100 descriptors are characterized by mean values falling within the same range observed for the random numbers [min = 0.466, max = 0.529]. Moreover, none of these 14 descriptors show a standard deviation within the range [min = 0.290, max = 0.300] found for our random numbers at $N = 100$. Furthermore, the number of the data points did not influence significantly the parameters of the probability density function. Descriptor values for any set of compounds are not random and in general are not uniformly distributed; for example, in a series with a prevailing number of compounds having high hydrophobicity (more compounds with high LogP values) the maximum of the probability density function characterizing the LogP descriptor will be shifted to the right and hence deviate significantly from uniform. Thus, from a statistical viewpoint random numbers and molecular descriptors behave quite differently; they are not interchangeable, and tests such as mode 1 are irrelevant.

As an additional study of equivalency, two data sets were compared - one consisted of 946 CODESSA PRO- calculated molecular descriptors and one of 946 random numbers. A forward stepwise multiple regression analysis on data sets consisted of 10, 20, 30, and 40 data points from the set of vapor pressure data for one parameter correlations gave the changes in R^2 shown in Table 7. The R^2 values obtained for the random numbers data set decrease rapidly with the number of data points in the set, while the R^2 values for "real" descriptors remain relatively constant. The conclusion is obvious: molecular descriptors are not random numbers. The descriptors encode features of molecular structure which can influence and control observed biological or physicochemical effects. While the interpretation of descriptors is not always clear or unique, in most cases they encode useful information — how successfully this information will be extracted depends on the skill of the researcher.

Table 7. Best R^2 Obtained for Data Sets of "Real" Descriptors and Random Numbers

number of cases	best R^2 (molecular descriptors)	best R^2 (random numbers)
10	0.91	0.65
20	0.87	0.36
30	0.86	0.14
40	0.83	0.05

4. CONCLUSIONS

QSPRs using large descriptor pools have been criticized for their increased sensitivity to chance correlations. However, as shown here, the possibility of chance correlations can be minimized to be negligible by using appropriate procedures. Crucially, the collinearity of natural descriptor scales should be strictly controlled during the forward selection processing. The criteria used in BMLR procedure have proven to be sufficient for the elimination of chance

correlations due to nonorthogonality of the scales. Second, the larger number of data points in the set gives additional guarantees for avoiding the chance correlations.

While tests with randomly generated scales might have possible significance, in such cases the size of the space generated by these random scales must be compatible with the size of the actual descriptor space.

Supporting Information Available: The orthogonality problem and SM Tables 1 and 2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Topliss, J. G.; Edwards, R. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(2) Topliss, J. G.; Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068.

(3) Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relat.* **1993**, *12*, 137–145.

(4) Doweyko, A. M. QSAR: dead or alive. *J. Comput.-Aided. Mol. Des.* **2008**, *22*, 81–89.

(5) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

(6) Katritzky, A. R.; Slavov, S.; Dobchev, D.; Karelson, M. Rapid QSPR model development technique for prediction of vapor pressure of organic compounds. *Comput. Chem. Eng.* **2007**, *31*, 1123–1130.

(7) Katritzky, A. R.; Fara, D.; Karelson, M. QSPR of 3-aryloxazolidin-2-one antibacterials. *Bioorg. Med. Chem.* **2004**, *12*, 3027–3035.

(8) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points With Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400–10407.

(9) Matsumoto, M.; Nishimura, T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Mod. Comput. Sim.* **1998**, *8*, 3–30.

(10) Statsoft STATISTICA v6.0. www.statsoft.com (accessed August 6, 2008).

(11) CODESSA PRO 1.0 RC2, CODESSA PRO QSPR/QSAR software. www.codessa-pro.com (accessed September 22, 2008).

CI8002073