y-Randomization and Its Variants in QSPR/QSAR

Christoph Rücker,*[,‡] Gerta Rücker,[§] and Markus Meringer[||]

Biozentrum, University of Basel, 4056 Basel, Switzerland, Institute of Medical Biometry and Medical Informatics, University of Freiburg, 79104 Freiburg, Germany, and Department of Medicinal Chemistry, Kiadis B.V., 9747 Groningen, The Netherlands

Received May 8, 2007

y-Randomization is a tool used in validation of QSPR/QSAR models, whereby the performance of the original model in data description (r^2) is compared to that of models built for permuted (randomly shuffled) response, based on the original descriptor pool and the original model building procedure. We compared y-randomization and several variants thereof, using original response, permuted response, or random number pseudoresponse and original descriptors or random number pseudodescriptors, in the typical setting of multilinear regression (MLR) with descriptor selection. For each combination of number of observations (compounds), number of descriptors in the final model, and number of descriptors in the pool to select from, computer experiments using the same descriptor selection method result in two different mean highest random r^2 values. A lower one is produced by y-randomization or a variant likewise based on the original descriptors, while a higher one is obtained from variants that use random number pseudodescriptors. The difference is due to the intercorrelation of real descriptors in the pool. We propose to compare an original model's r^2 to both of these whenever possible. The meaning of the three possible outcomes of such a double test is discussed. Often y-randomization is not available to a potential user of a model, due to the values of all descriptors in the pool for all compounds not being published. In such cases random number experiments as proposed here are still possible. The test was applied to several recently published MLR QSAR equations, and cases of failure were identified. Some progress also is reported toward the aim of obtaining the mean highest r^2 of random pseudomodels by calculation rather than by tedious multiple simulations on random number variables.

INTRODUCTION

Whenever in QSPR/QSAR model building a "best" combination of a few (m) descriptors is selected from a pool of many (M) descriptors in order to best fit given data, there is an enhanced risk of chance correlation, as was pointed out by Topliss et al. in the 1970s.[1,2] The risk is enhanced (compared to using a prespecified set of descriptors) due to the number of possible models considered,

$$^{M}C_{m} = M!/[m!(M - m)!]$$

which for increasing m and M quickly grows to astronomic orders of magnitude. Among so many combinations of m descriptors, a few are likely to fit the data reasonably well by chance, i.e., without having a true connection to the response variable. Descriptor selection procedures, led by a simple measure of fit, are blind with respect to presence or absence of such a true connection.

In the 1970s, few molecular descriptors were available to select from, so that in QSPR/QSAR work chance correlation was a minor risk. At present, with several computer programs available that routinely calculate hundreds or even thousands of molecular descriptors and then automatically select a "best" subset from these,[3−5] the problem (now called selection bias[6]) has become urgent. Even worse, as more and more molecular descriptors are being developed, each particular researcher's descriptor pool must be considered a more or less arbitrary subset of the set of all descriptors that are or will become available in the future.

In 1980 Rencher and Pun quantified the problem by computer simulations, calculating average and upper 95th percentile of r^2 for describing n data points by a best MLR model containing m descriptors selected from a pool of M random number pseudodescriptors by a stepwise selection procedure ($n \leq 60$, $m \leq 10$, $M \leq 40$).[7] To enable interpolations for other (n,m,M) tuples they described their experimental results as a highly nonlinear equation. The results were paraphrased by Wold:[8] *"... if we have sufficiently many structure descriptor variables to select from we can make a model fit data very closely even with few terms, provided that they are selected according to their apparent contribution to the fit. And this even if the variables we choose from are completely random and have nothing whatsoever to do with the current problem!"*

In a similar manner Livingstone and Salt performed computer experiments of fitting random number response by random number descriptors for various (n,m,M) tuples ($n \leq 100$, $m \leq 8$, $M \leq 100$). They emphasized that normal tabulated F values are inadequate in the case of descriptor selection and concentrated on modifying F appropriately.[6a] Recently these authors, in a remarkable effort based on huge

* Corresponding author phone: +049 761 484079; fax: +049 761 2036680; e-mail: christoph.ruecker@uni-bayreuth.de.
‡ University of Basel.
§ University of Freiburg.
|| Kiadis B.V.

10.1021/ci700157b CCC: $37.00
Published on Web 09/20/2007

computer power, extended the range up to $n \leq 100$, $m \leq 10$, $M \leq 150$, even using all-subsets regression.[6b] For assessing our MLR models[9−11] and others from the literature, however, refs 6a,b and 7 are of little help due to their still limited (n,m,M) range. This range, on the other hand, is not likely to be sufficiently extended in the near future, due to the huge amount of calculations necessitated by the number ${}^{M}C_{m}$.

As a minimum requirement, a model useful for prediction or understanding should describe the given data better than chance would do, that is, it should be statistically significant. Therefore we became interested in the following question: *How well could our response data be fitted by pure chance, i.e. by selecting the "best" combination of a few (m) out of many (M) random number pseudodescriptors?* (question 1). In our opinion this is not only a legitimate but also a necessary question to be asked; it is in accord with accepted reasoning in statistics, e.g., in the theory of statistical tests.

A popular tool used by researchers to protect themselves against the risk of chance correlation has been y-randomization (also called y-scrambling[12] or response randomization[13]), a method said to be "*probably the most powerful validation procedure*".[12] By validation developers try to convince themselves of a model's properties such as statistical significance, robustness, predictive ability, etc.[13,14] While other important validation methods such as cross-validation and training set/test set partitioning were discussed in detail recently,[14−17] y-randomization is often applied but did not attract much attention itself.[18] While it is mentioned, without any details given, in the books written by Harrell[19] and by Manly,[20] we did not find it in the potentially relevant book by Miller.[21]

y-Randomization was used early, e.g., by Klopman and Kalos.[22] It was nicely described in a paper by Wold and Eriksson[23] (and similarly by Karki and Kulkarni[24]): *"The first of the four tools is based on repetitive randomization of the response data (Y) of N compounds in the training set. Thus, a random number generator is used to allocate the integers between 1 and N to sequences of N numbers. In each cycle, the resulting arrangement of random integers is employed in order to reorder the Y data − leaving the X data intact − and then the full data analysis is carried out on these scrambled data. Every run will yield estimates of R^2 and Q^2, which are recorded. If in each case the scrambled data give much lower R^2 and Q^2 values than the original data, then one can feel confident about the relevance of the "real" QSAR model."*

The authors did not give a reference, nor did they provide a mathematical justification for the procedure.

Note in the quotation the important phrase *"and then the full data analysis is carried out"*. This includes descriptor selection starting from the full pool of initial descriptors for each y-randomized run. Occasionally in the literature y-randomized procedures are encountered that do not include descriptor selection, instead they use the m descriptors of the final model to describe the scrambled y data, thus at all ignoring the problem. This misunderstanding of y-randomization results in a very poor fit for the random models, giving an extremely overoptimistic impression of the original model. In the Results section this is illustrated in detail.

Since a particular permutation of y values may be close to the original arrangement, a single or a few out of many y-permutation runs may result in a rather high fit without saying that the model under scrutiny is spurious.[23,25] Therefore it is occasionally difficult to decide from the outcome of y-randomization whether or not a model has passed the test. A quantitative evaluation of the test result, in the framework of standard statistical hypothesis testing, was given recently.[26]

By the above, y-randomization is an attempt to observe the action of chance in fitting given data. This is done by repeatedly and deliberately destroying the connection between response variable y and independent variables x (in QSPR/QSAR: molecular descriptors) by randomly permuting the y data, leaving all x data untouched, and performing the whole model building procedure as it would be done for real y data. Thus, y-randomization may be considered a special case of permutation tests.[27] y-Randomization asks and answers the following question: *How well could random scramblings of my response data be fitted by selecting the "best" combination of m out of my M descriptors?* (question 2). Note that this question differs from question 1 above in two respects. First, question 1 asks for the original y data, while question 2 asks for a random scrambling of y data. Second, in question 1 random number pseudodescriptors are considered, in question 2 the original descriptors. Further analogous questions can be raised, such as the following: *How well could random numbers be fitted by selecting the "best" combination of m out of my M descriptors?* (question 3); *How well could random numbers be fitted by selecting the "best" combination of m out of M random number pseudodescriptors?* (question 4); and *How well could random scramblings of my response data be fitted by selecting the "best" combination of m out of M random number pseudodescriptors?* (question 5). There is no a priori reason to expect the answers of these five questions to be identical. We therefore decided to approach these questions in a comparative manner using computer experiments, concentrating on (n,m,M) tuples from recent QSPR/QSAR studies. In the first part of the present article we report the results, whereby a picture was obtained of which r^2 value to expect for given (n,m,M) by the action of chance alone. Such numbers obviously can be used as warning thresholds, in that a statistically significant MLR model has to considerably surpass them. In the second part we use such information to assess some recently published MLR QSAR models. In the third part a misunderstanding of the y-randomization procedure is clarified. In the fourth part, in order to obviate tedious multiple computer runs on random number variables, we introduce a small program to calculate lower and upper bounds of and an approximation for the mean highest random r^2 (mhr r^2) expected by selection from a pool of random pseudodescriptors in a (n,m,M) MLR situation.

For simplicity and transparency, here y-randomization and its variants are considered in the context of multilinear regression (MLR) only, though y-randomization was used in connection with many other QSAR methods also.[3,14,26,28,29]

METHODS

Data Sets. We extracted from the literature data sets containing, along with the final MLR model and the identities and target activities of n compounds, the values of all M descriptors in the pool for all compounds. While such data sets are rare, we found the following three.

Kier and Hall reported an MLR model of the hallucinogenic activity of 23 substituted amphetamines, where three descriptors were selected from a pool of 18 (data set 1).[30]

The so-called Selwood data set consists of 16 antifilarial antimycin analogs together with their activities and the values of 53 descriptors (data set 2).[31]

Prabhakar et al. described the aldose reductase inhibitory activity of 48 flavones by MLR models containing between 3 and 7 descriptors from a pool of 158 (data set 3).[32]

Another few data sets including values of all *M* descriptors for all *n* compounds were available to us from our previous work. Thus, we recently proposed a MLR model for the binding affinity of 144 PPARγ ligands, containing 10 descriptors selected from a 230-descriptor pool (data set 4).[11] In the same work, the gene transactivation activity of 150 such ligands was described by a MLR model on 14 out of 229 descriptors (data set 5).

Earlier, we described the boiling points of 507 C_1-C_4 haloalkanes by MLR models using 6 or 7 descriptors from a pool of 249 (data set 6)[9] and the boiling points of 82 C_1-C_4 fluoroalkanes using 6 or 7 descriptors out of 209 (data set 7).[10]

For the following data sets values of all descriptors in the pool are not available.

The COX-2 inhibitor activities of 24 terphenyls (data set 8) and of 15 4,5-diphenyl-2-trifluoromethylimidazoles (data set 9) were modeled using 4 and 3 descriptors, respectively, by Hansch et al.[33] The descriptor pool consisted of at least 14 variables in both cases.

The antibacterial activities of 60 oxazolidinones (data set 10) were treated by Karki and Kulkarni[24] and later by Katritzky et al. who selected from a pool of no less than 1627 descriptors.[5]

The antiplasmodial activities of 16 cinnamic acid derivatives (data set 11) were described as a 3-descriptor equation by Gupta et al.[34] The same research group treated binding to PPARγ of 16 2-benzoylaminobenzoic acids (data set 12) and gene transactivation effected thereby (data set 13) as well as binding of a subset of these to PPARα (data set 14).[35]

Prabhakar et al. treated the antimycobacterial activity of two series of functionalized alkenols (data sets 15 and 16).[36]

Descriptor Selection Procedure. In MLR, for a given data set consisting of a target variable and *M* descriptors for *n* compounds, that combination of *m* descriptors ($m < M$) is sought that results in the best fit among all possible *m*-descriptor models. Best fit means highest r^2 or equivalently highest *F*, since these quantities are monotonically related for constant *n* and *m*. Running through *all* combinations usually is unfeasible (too time-consuming); therefore, several approximate methods have been proposed for this purpose (forward inclusion, backward elimination, stepwise methods, genetic algorithms, etc.[21,37]), but none is guaranteed to always find the very best combination. The "best" model found for a given data set may differ from method to method. So a real QSAR model should be compared to pseudomodels based on randomization preferably using the same descriptor selection procedure.

We restricted ourselves to the step-up procedure as implemented in MOLGEN-QSPR.[9,10] This procedure calculates all 1-descriptor models, combines the best *l* of these each with all other descriptors one by one, takes the best *l* of these 2-descriptor models to combine each with a third descriptor, and so on. The parameter *l* was set to 1000 in this work. Reasons for choosing this method included the following:

(i) The step-up method, a kind of forward selection, despite its apparent simplicity resulted in high-r^2 models in our previous work; it proved even not inferior to the genetic algorithm provided in the commercial MOE package.[10] Additionally, we tested the method by comparing its results to those obtained by Rencher and Pun by their stepwise method for 91 (*n*,*m*,*M*) tuples between (5,2,5) and (60,10,40).[7] Our method consistently, i.e., for each of the 91 tuples, resulted in slightly higher r^2 values.

(ii) For smaller *m* and *M*, this method is equivalent to an exhaustive (all-subset) search. This is the case in all experiments reported below in Tables 1 and 7, and all lines except one in Tables 2 and 5.

(iii) Even if we had an "ideal" descriptor selection method at our hands (always finding that combination of descriptors resulting in the highest possible r^2), such a method would be likely to find a higher r^2 model both for the original data matrix and for each randomized data matrix, so that the result of comparison between original model and models originating from randomized data would be expected to be similar as with a real (nonideal) descriptor selection method.

(iv) We planned to assess by randomization experiments published models from the literature. Most of these were obtained using common descriptor selection methods such as forward or stepwise selection, i.e., not using an ideal method. Therefore it would be unfair to the authors of published work to use an ideal descriptor selection procedure in the randomization experiments.

Planned application to published models further necessitated the use of r^2 as the measure of fit.

Random Numbers. Pseudorandom integers uniformly distributed between 0 and 32767 (2^{15}-1) were generated by the C function rand( ), using the system time as random

Table 1. Experimental mhr r^2 Values and *Standard Deviations* for the Tuple (23,3,18) Belonging to Data Sets 1, 1A, and 1B[a]

n	*m*	*M*	*it*	mode 1 y vs rx	mode 2 py vs x	mode 3 ry vs x	mode 4 ry vs rx	mode 5 py vs rx
				Data Set 1				
23	3	18	25	0.4081	0.3290	0.3254	0.4306	0.4572
				0.0930	*0.1293*	*0.0890*	*0.1133*	*0.1048*
23	3	18	25	0.4459	0.2850	0.3270	0.4016	0.4021
				0.0938	*0.1135*	*0.0974*	*0.0654*	*0.0880*
23	3	18	25	0.4280	0.3188	0.3233	0.4665	0.4485
				0.1178	*0.1125*	*0.0943*	*0.1142*	*0.1222*
23	3	18	25	0.4373	0.3205	0.3012	0.4155	0.4307
				0.0956	*0.0999*	*0.0936*	*0.1016*	*0.1335*
23	3	18	250	0.4352	0.3194	0.3166	0.4406	0.4159
				0.1041	*0.1368*	*0.1006*	*0.1066*	*0.1047*
23	3	18	2500	0.4280	0.3185	0.3158	0.4335	0.4299
				0.1054	*0.1250*	*0.1089*	*0.1087*	*0.1040*
23	3	18	25000	0.4291	0.3181	0.3147	0.4323	0.4312
				0.1048	*0.1264*	*0.1099*	*0.1083*	*0.1058*
				Data Set 1A				
23	3	18	2500	0.4291	0.2951	0.2919	0.4311	0.4334
				0.1079	*0.1228*	*0.1094*	*0.1079*	*0.1055*
				Data Set 1B				
23	3	18	2500	0.4278	0.4294	0.4305	0.4328	0.4298
				0.1060	*0.1052*	*0.1093*	*0.1101*	*0.1054*

[a] For comparison, r^2 of the original model is 0.846.

Table 2. Experimental mhr r^2 Values and *Standard Deviations* for (n,m,M) Tuples Belonging to Data Set 2[a]

n	m	M	mode 1 y vs rx	mode 2 py vs x	mode 3 ry vs x	mode 4 ry vs rx	mode 5 py vs rx	orig r^2
16	1	53	0.3770	0.3052	0.3043	0.3822	0.3803	(0.49)
			0.0985	*0.1017*	*0.0956*	*0.0978*	*0.1008*	
16	2	53	0.6223	0.5287	0.5192	0.6242	0.6201	(0.74)
			0.0847	*0.1110*	*0.1073*	*0.0940*	*0.0940*	
16	3	53	0.7885	0.6756	0.6730	0.7899	0.7866	(0.81)
			0.0586	*0.1008*	*0.0970*	*0.0635*	*0.0663*	
16	1	23	0.3094	0.2980	0.2819	0.3181	0.3183	0.49
			0.1053	*0.1161*	*0.0984*	*0.1106*	*0.1046*	
16	2	23	0.5077	0.4710	0.4785	0.5038	0.5047	0.74
			0.1112	*0.1143*	*0.1127*	*0.1208*	*0.1170*	
16	3	23	0.6445	0.6230	0.6184	0.6378	0.6480	0.81
			0.1118	*0.1136*	*0.1088*	*0.1085*	*0.1079*	
16	1	10	0.2374	0.2271	0.2204	0.2343	0.2404	
			0.1130	*0.0945*	*0.0973*	*0.1079*	*0.1125*	
16	2	10	0.3694	0.3789	0.3592	0.3810	0.3777	
			0.1170	*0.1288*	*0.1316*	*0.1390*	*0.1367*	
16	3	10	0.4745	0.4624	0.4589	0.4827	0.4676	
			0.1446	*0.1327*	*0.1416*	*0.1407*	*0.1355*	

[a] $it = 250$ throughout. For comparison, in the last column r^2 of the original models are given.

argument for srand( ) in order to obtain a new sequence of pseudorandom numbers in every run.[38] For use as random pseudodescriptors these numbers were taken as such; for use as random pseudoresponse they were scaled to the range of original response data. To obtain random permutations, the C++ function random_shuffle( ) was used, again based on pseudorandom integers obtained from rand( ) and seeded by srand( ) and the system time.

Randomization Experiments. For each data set, for its particular triple (n,m,M) a "best" MLR pseudomodel was established using the MOLGEN-QSPR step-up descriptor selection procedure, either after replacing the target variable by a random permutation of the given values, or after replacing original variables by random number pseudovariables, in five different modes that correspond to the five questions raised above:

(i) mode 1, original target variable, descriptors replaced by M pseudodescriptors made of random numbers (y vs rx);

(ii) mode 2, target variable randomly permuted, M original descriptors (py vs x, y-randomization);

(iii) mode 3, target variable replaced by random numbers, M original descriptors (ry vs x);

(iv) mode 4, target variable replaced by random numbers, descriptors replaced by M pseudodescriptors made of random numbers (ry vs rx); and

(v) mode 5, target variable randomly permuted, descriptors replaced by M pseudodescriptors made of random numbers (py vs rx).

In the shorthand notation given, p stands for permuted, and r stands for random number. For each mode, this procedure was repeated it times ($it \geq 25$), each time using a fresh set of random numbers. In each repetition the highest r^2 value obtained by descriptor selection was recorded (highest random r^2), and the mean highest random (mhr) r^2 and its standard deviation were calculated by averaging over it repetitions.

RESULTS

1. Computer Simulations of Chance Correlation. The results of our randomization experiments are reported in Tables 1−7, where for several (n,m,M) combinations from literature data sets experimental mhr r^2 values (upright) are given together with the corresponding standard deviations (*italic*), separately for the five modes.

Table 1 contains results for data set 1. Comparison of the first four lines shows the scatter due to random. In the next three lines the number of repetitions it was varied. Neither mhr r^2 nor standard deviations differ substantially between $it = 25$ and $it = 25\,000$. We conclude from this result that for our purposes $it = 25$ is sufficient, though of course higher it values are desirable.

The foremost result, apparent in all lines of Table 1 (and in Tables 2−4 as well, see below) is the following: mhr r^2 values obtained from mode 2 and from mode 3 agree (within the limits of random scatter) and are lower by some margin than those from modes 1, 4, and 5, which again agree.

Table 3. Experimental mhr r^2 Values and *Standard Deviations* for (n,m,M) Tuples Belonging to Data Set 3[a]

n	m	M	mode 1 y vs rx	mode 2 py vs x	mode 3 ry vs x	mode 4 ry vs rx	mode 5 py vs rx	orig r^2
48	3	158	0.4158	0.2909	0.2655	0.4120	0.4171	0.608
			0.0537	*0.0831*	*0.0662*	*0.0640*	*0.0659*	
48	4	158	0.4866	0.3472	0.3620	0.4995	0.5011	0.682
			0.0411	*0.0694*	*0.0893*	*0.0481*	*0.0526*	
48	5	158	0.5598	0.3876	0.3993	0.5814	0.5833	0.667
			0.0447	*0.0756*	*0.1049*	*0.0491*	*0.0487*	
48	6	158	0.6465	0.4447	0.4490	0.6724	0.6676	0.752
			0.0469	*0.0921*	*0.0926*	*0.0526*	*0.0435*	
48	7	158	0.7188	0.4586	0.4860	0.7159	0.7148	0.778
			0.0441	*0.0812*	*0.0764*	*0.0436*	*0.0366*	
32	3	158	0.5905	0.3663	0.3879	0.5891	0.5924	
			0.0518	*0.0841*	*0.0691*	*0.0658*	*0.0522*	
32	4	158	0.6834	0.4596	0.4707	0.6897	0.6844	
			0.0408	*0.0852*	*0.1135*	*0.0556*	*0.0458*	
32	5	158	0.7722	0.5767	0.5453	0.7862	0.7866	
			0.0265	*0.0761*	*0.1135*	*0.0364*	*0.0507*	
32	6	158	0.8418	0.6047	0.6117	0.8469	0.8418	
			0.0219	*0.1038*	*0.0917*	*0.0298*	*0.0353*	
32	7	158	0.8975	0.6360	0.6776	0.8911	0.8875	
			0.0214	*0.0917*	*0.0845*	*0.0228*	*0.0145*	

[a] $it = 25$ throughout. For comparison, in the last column r^2 of the original models are given.

Table 4. Experimental mhr r^2 Values and *Standard Deviations* for (n,m,M) Tuples Belonging to Data Sets 4−7[a]

n	m	M	mode 1 y vs rx	mode 2 py vs x	mode 3 ry vs x	mode 4 ry vs rx	mode 5 py vs rx	orig r^2
					Data Set 4			
144	10	230	0.3826	0.3060	0.2865	0.3839	0.3899	0.7938
			0.0348	*0.0430*	*0.0395*	*0.0431*	*0.0356*	
129	10	230	0.4358	0.3223	0.3276	0.4308	0.4403	0.7909
			0.0316	*0.0458*	*0.0533*	*0.0366*	*0.0333*	
					Data Set 5			
150	14	229	0.4524	0.3506	0.3434	0.4607	0.4653	0.6487
			0.0400	*0.0457*	*0.0336*	*0.0271*	*0.0335*	
					Data Set 6			
507	6	249	0.0799	0.0542	0.0552	0.0798	0.0775	0.9879
			0.0137	*0.0103*	*0.0112*	*0.0110*	*0.0130*	
507	7	249	0.0876	0.0599	0.0624	0.0915	0.0882	0.9888
			0.0107	*0.0095*	*0.0125*	*0.0120*	*0.0081*	
					Data Set 7			
82	6	209	0.4168	0.2835	0.2753	0.4400	0.4335	0.9845
			0.0369	*0.0529*	*0.0688*	*0.0320*	*0.0364*	
82	7	209	0.4745	0.2977	0.2929	0.4859	0.4773	0.9872
			0.0386	*0.0386*	*0.0491*	*0.0409*	*0.0321*	

[a] $it = 25$ throughout. For comparison, in the last column r^2 of the original models are given.

Table 5. Experimental mhr r^2 Values and *Standard Deviations* for (*n*,*m*,*M*) Tuples Belonging to Data Sets 8 and 9[e]

n	*m*	*M*	mode 1 y vs rx	mode 2 py vs x	mode 3 ry vs x	mode 4 ry vs rx	mode 5 py vs rx	orig r^2
				Data Set 8				
24	4	14[a]	0.4358	(0.3070)	(0.3083)	0.4251	0.4255	0.909
			0.1158	*(0.1308)*	*(0.1252)*	*0.1104*	*0.1019*	
24	4	14[b]	0.4254	(0.4297)	(0.4204)	0.4291	0.4350	
			0.1168	*(0.1142)*	*(0.1217)*	*0.1145*	*0.1087*	
27	4	14	0.3841			0.3836	0.3856	(0.661)
			0.1007			*0.1118*	*0.1060*	
27	4	25	0.4889			0.5028	0.5016	(0.661)
			0.0879			*0.0960*	*0.0964*	
				Data Set 9				
15	3	14[c]	0.5767	(0.4211)	(0.4191)	0.5660	0.5783	0.885
			0.1222	*(0.1590)*	*(0.1561)*	*0.1407*	*0.1347*	
15	3	14[d]	0.5634	(0.5855	(0.5857)	0.5727	0.5763	
			0.1277	*(0.1248)*	*(0.1271)*	*0.1287*	*0.1280*	
17	3	14	0.5270			0.5018	0.5134	(0.777)
			0.1333			*0.1368*	*0.1342*	
17	3	25	0.6377			0.6350	0.6415	(0.777)
			0.0982			*0.1019*	*0.0929*	

[a] Four original descriptors and 10 highly intercorrelated topological indices. [b] Four original descriptors and 10 random pseudodescriptors. [c] Three original descriptors and 11 highly intercorrelated topological indices. [d] Three original descriptors and 11 random pseudodescriptors. [e] *it* = 250 throughout. For comparison, in the last column r^2 of the original models are given.

Table 6. Experimental mhr r^2 Values and *Standard Deviations* for (*n*,*m*,*M*) Tuples Belonging to Data Set 10[b]

n	*m*	*M*	*it*	mode 1 y vs rx	mode 4 ry vs rx	mode 5 py vs rx	orig r^2
50	3	34	250	0.2607	0.2539	0.2567	0.603
				0.0546	*0.0633*	*0.0612*	
50	4	34	250	0.3210	0.3204	0.3107	0.651
				0.0684	*0.0740*	*0.0647*	
50	6	34	250	0.3999	0.4004	0.3961	0.732
				0.0727	*0.0714*	*0.0773*	
50	4	10	250	0.1644	0.1634	0.1669	
				0.0726	*0.0706*	*0.0712*	
60	7	1627	25	0.8188[a]	0.8181	0.8146	0.820
				0.0183[a]	*0.0172*	*0.0176*	
60	7	888	25	0.7772	0.7733	0.7684	0.795
				0.0245	*0.0207*	*0.0221*	
60	7	739	25	0.7635	0.7560	0.7641	0.731
				0.0244	*0.0234*	*0.0268*	
50	6	1627	25	0.8350	0.8319	0.8335	0.809
				0.0157	*0.0233*	*0.0151*	

[a] Another series (*it* = 50) yielded 0.8128, standard deviation *0.0151*. [b] For comparison, in the last column r^2 of the original models are given.

The difference is due to the original descriptors (used in modes 2 and 3 only) being intercorrelated. Rencher and Pun already had observed lower mhr r^2 values when using intercorrelated instead of noncorrelated random number pseudodescriptors.[7] In fact, the 18 descriptors in data set 1, six connectivity χ indices along with their squares and reciprocals, are highly intercorrelated. To test the influence of descriptor intercorrelation, we replaced in data set 1 the original descriptors by either 18 highly intercorrelated topological indices (data set 1A) or by 18 random pseudodescriptors (data set 1B) and repeated the whole series of experiments. It was expected that the result for data set 1A would be similar to that of data set 1, while in data set 1B the difference between the modes should vanish. This is exactly what happened (see the last lines of Table 1). Similar influence of descriptor intercorrelation was observed for data sets 8 and 9, see below.

Table 7. Experimental mhr r^2 Values and *Standard Deviations* for the (*n*,*m*,*M*) Tuples Belonging to Data Sets 11−16[a]

n	*m*	*M*	mode 1 y vs rx	mode 4 ry vs rx	mode 5 py vs rx	orig r^2
			Data Set 11			
16	3	33	0.7079	0.7191	0.7156	0.689
			0.0808	*0.0862*	*0.0862*	
			Data Set 12			
16	3	32	0.7051	0.7234	0.7150	0.808
			0.0852	*0.0881*	*0.0910*	
			Data Set 13			
15	3	32	0.7443	0.7425	0.7555	0.750
			0.0810	*0.0832*	*0.0756*	
			Data Set 14			
8	1	32	0.5838	0.6117	0.5840	0.738
			0.1289	*0.1258*	*0.1368*	
			Data Set 15			
11	2	96	0.8644	0.8665	0.8690	0.748
			0.0490	*0.0462*	*0.0480*	
			Data Set 16			
11	2	96	0.8482	0.8571	0.8503	0.733
			0.0491	*0.0486*	*0.0474*	

[a] *it* = 250. For comparison, in the last column r^2 of the original models are given.

Results for data set 2 ($n = 16$) are shown in Table 2, all obtained with *it* = 250. In the original paper the initial set of 53 highly intercorrelated descriptors was narrowed down to 23 weakly intercorrelated ones by removing one descriptor from each pair intercorrelated higher than $r = 0.75$. Out of these 23, 10 descriptors were selected according to their correlation with the target variable, and from these 1-, 2-, and 3-descriptor models were selected.[31] We therefore treated all such (16,*m*,*M*) triples.

The first thing to notice in Table 2 is the magnitude of the entries. Thus, for 16 observations (compounds), selection of the best combination of 3 out of 53 descriptors leads to highest random $r^2 = 0.79$ on average even if all descriptors are purely random. This is true for the original response data (mode 1), for random pseudoresponse (mode 4), and for randomly permuted response (mode 5). With the 53 original descriptors from data set 2 the corresponding mhr r^2 is still 0.67, both for permuted response and for random number pseudoresponse (modes 2 and 3). Obviously, selection bias is everything but negligible.

In Table 2, mhr r^2 increases with increasing *m* for constant *n* and *M* and with increasing *M* for constant *n* and *m*, as it should. The mhr r^2 difference between modes 2/3 and modes 1/4/5 is large in the set of 53 highly intercorrelated descriptors (14−19%) and smaller in the subset of 23 weakly intercorrelated descriptors (3−7%), confirming the earlier results and our view on the origin of this difference.

Table 3 shows our results for data set 3 ($n = 48$, $M = 158$, *it* = 25 throughout). In the original paper 3- through 7-descriptor models were given; we therefore treated the corresponding (48,*m*,158) triples.

To test the influence of *n* for constant *m* and *M*, we eliminated the last 16 compounds from data set 3 and repeated all experiments for the remaining 32. As expected, all mhr r^2 values increased with this decrease in *n*.

Table 4 shows the results for data sets 4−7, $it = 25$ throughout. For data set 4, in the original paper,[11] a 129-compound subset (training set) of the complete 144-compound set was treated as well, and accordingly here experiments for (129,10,230) are included. Compared to the previous data sets, m and M are considerably increased here, but their influence is counterbalanced by increased n, so that for data sets 4 and 5 the resulting mhr r^2 values are in the low to middle range again. For data set 6, high $n = 507$ together with low $m = 6$ or 7 cause mhr r^2 to drop to very low numbers even for rather high $M = 249$.

We summarize the content of Tables 1−4 as follows:

(i) mhr r^2 values increase with increasing m and M and with decreasing n.

(ii) Permuted response values or random number pseudoresponse are fitted equally well on average by best combinations of the original descriptors (mode 2 and mode 3).

(iii) Original response values, random number pseudoresponse, or randomly permuted response are fitted equally well on average by best combinations of random pseudodescriptors (modes 1, 4, and 5).

(iv) Best combinations of m original descriptors (modes 2 and 3) are less successful on average in establishing chance correlations than best combinations of m random pseudodescriptors (modes 1, 4, and 5, compare in particular mode 3 to mode 4 and mode 2 to mode 5). This is due to intercorrelation usually found among real descriptors.

Thus, for each original model, randomization experiments yield two different mhr r^2 values for possible comparison with the original r^2, a lower one from experiments using the original descriptors such as y-randomization, and a higher one from experiments using random number pseudodescriptors. The meaning of these two numbers and the question of which one should be used for comparison are discussed in the Discussion section. In the following section we compare each original model's r^2 to both corresponding mhr r^2 values, whenever possible.

2. Application to Published QSAR Equations. We approximate the distribution of highest random r^2 values for a given (n,m,M) tuple by a normal distribution. The difference between r^2 of an original MLR model and mhr r^2 then should roughly be ≥ 2.3 standard deviations (SD) for significance on the 1% level, ≥ 3 SD for the 0.1% level, etc.

Data Set 1. The published r^2 of the original 3-descriptor model[30] (0.846) is higher than the mode 1 mhr r^2 for 23 compounds and 3 out of 18 descriptors (0.4291, SD = 0.1048, Table 1) by 4.0 SD, and higher than the mode 2 (y-randomization) mhr r^2 (0.3181, SD = 0.1264, Table 1) by 4.2 SD. Averaging over modes 1/4/5 or modes 2/3 leads to similar numbers. The original eq 1 therefore is safe in the sense that it fits the data significantly better than chance correlations.[39]

Data Set 2. In the original paper 1-, 2-, and 3-descriptor equations are given having $r^2 = 0.49$, 0.74, and 0.81, respectively.[31] Had these equations been obtained by descriptor selection from the original pool of 53 descriptors, the differences between r^2 and mhr r^2 would be 1.15, 1.39, and 0.37 SD, respectively (mode 1, Table 2), or 1.81, 1.90, and 1.33 SD (mode 2). Averaging over modes 1/4/5 or over modes 2/3 leads to essentially the same numbers. The equations therefore could not be considered significant by either test.

However, the original equations were obtained by descriptor selection from the pool of 23. The r^2 distances from mhr r^2 (mode 1, Table 2) are 1.72, 2.09, and 1.48 SD, respectively, so that the 2-descriptor equation, if any, is close to what may be considered significant in the more demanding test, in complete agreement with the results of Livingstone and Salt.[6a] In y-randomization (mode 2) the distances are 1.65, 2.35, and 1.65 SD, respectively, leading to the same conclusion.

Data Set 3. In the original paper MLR models containing 3, 4, and 5 descriptors are given with $r^2 = 0.608$, 0.682, and 0.667, respectively.[32] These were found by descriptor selection that was restricted by some filters, from a pool of 158 descriptors. The real r^2 values are higher than the respective mode 1 mhr r^2 by 3.58, 4.75, and 2.40 SD (Table 3). y-Randomization (mode 2) leads to distances between original r^2 and mhr r^2 of 3.82, 4.82, and 3.70 SD.

Thus while the first two equations are safe by both tests, the 5-descriptor equation is safe by the y-randomization test, but a boundary case according to the more demanding random number test. The original equations were obtained in a procedure that prohibited simultaneous appearance in the same model of descriptors intercorrelated by $r > 0.3$, which in a similar manner as in data set 2 may have efficiently diminished the effective number of descriptors to select from.

Finally, in the original paper a 6- and a 7-descriptor model are proposed ($r^2 = 0.752$ and 0.778) that formally were obtained by descriptor selection from a 26-descriptor pool. This pool contained all descriptors appearing in a set of models that emerged by descriptor selection from the pool of 158. Therefore random experiments using $M = 26$ would be too optimistic here, and experiments using $M = 158$ resulted in distances of r^2 from mode 1 mhr r^2 of 2.25 and 1.34 SD for the 6- and the 7-descriptor model, respectively, while y-randomization (mode 2) resulted in distances of 3.34 and 3.93 SD. Therefore by y-randomization these models are safe, whereas by the more demanding random number test their significance is not beyond doubt.

Data Set 4. In the original paper r^2 of model m1 (binding of 144 PPARγ ligands) is given as 0.7938, which is more than 11 SD above mode 1 or mode 2 mhr r^2 (Table 4).[11] For the subset of 129 ligands, r^2 of model m2 is 0.7909, which likewise is more than 11 SD above mode 1 and more than 10 SD above mode 2 mhr r^2. Both models therefore are statistically significant.

Data Set 5. For gene transactivation induced by 150 PPARγ ligands, model m3 in the original paper has $r^2 = 0.6487$,[11] which is 4.9 SD above the mode 1 mhr r^2 (0.4524, Table 4) and 6.5 SD above the mode 2 mhr r^2 (0.3506), so that the original model is considered significant.

Data Set 6. For the boiling points of 507 haloalkanes, a MLR model of $r^2 = 0.9879$ was reported in the original paper, containing 6 descriptors that were selected from a pool of 249.[9] Comparison with the mode 1 and mode 2 mhr r^2 (0.0799 and 0.0542, Table 4) results in distances of 66 and 91 SD. For the 7-descriptor model the original r^2 is 0.9888, which is 84 and 98 SD above the mode 1 and mode 2 mhr r^2 for (507,7,249) (Table 4). Thus the statistical significance of both models is beyond any doubt.

Data Set 7. For the boiling points of 82 fluoroalkanes, the original MLR model containing 6 descriptors out of a

pool of 209 has $r^2 = 0.9845$,[10] which is 15 and 13 SD above mode 1 and mode 2 mhr r^2 (Table 4). For the 7-descriptor model, its $r^2 = 0.9872$ is 13 and 18 SD above mode 1 and mode 2 mhr r^2. Both models thus are statistically significant.

For the following data sets mode 2 and mode 3 simulations were impossible due to the original descriptor values being unavailable. Fortunately, the (minimum) numbers of descriptors in the pools were given, so that we were able to perform mode 1, mode 4, and mode 5 simulations.

Data Sets 8 and 9. A MLR QSAR equation for the COX-2 inhibitor activity of 24 substituted terphenyls (4 descriptors) was proposed by Hansch et al.[33] The COX-2 inhibitory activity of 15 substituted 4,5-diphenyl-2-trifluoromethylimidazoles (3 descriptors) was also modeled there. The descriptor pool consisted of at least 14 variables in both cases. In the paper numerical values are given for the 4 and 3 descriptors only that appear in the final models.

The original model for the terphenyls has $r^2 = 0.909$, more than 4 SD above the mode 1 mhr r^2 for (24,4,14) (0.43, Table 5), so that the original model seems significant at first sight. However, the compound set initially consisted of 27 terphenyls, of which three were excluded as outliers. From the data given,[33] the corresponding model for the 27-compound set has $r^2 = 0.661$, which is only 2.75 SD above the mode 1 mhr r^2 for (27,4,14) (0.3841, Table 5).

For the diphenylimidazoles the situation is similar. For the original model $r^2 = 0.885$ is 2.5 SD above mode 1 mhr r^2 (0.57, Table 5). However, two compounds had been excluded as outliers, and from the data given,[33] r^2 of the corresponding model for the 17-compound set can be calculated to be 0.7765, which is only 1.9 SD above mode 1 mhr r^2 for (17,3,14) (0.5270, Table 5).

The significance of both original models therefore is not beyond doubt, according to the random number test. Note that this conclusion is arrived at even applying the minimal $M = 14$ in both cases. Had we more realistically used $M = 25$ for the terphenyls (2 substituents, additional 11 substituent descriptors) or for the diphenylimidazoles (substituent position 2, 3, or 4 differentiated, resulting in additional descriptors), the significance of the original models would appear even more questionable, see the remaining mode 1 results in Table 5. In light of these results, a significance check for the other QSAR equations given in ref 33 seems highly desirable.

Data sets 8 and 9 provided another opportunity to test our understanding of the difference between mode 1/4/5 and mode 2/3 results. Thus, the results just mentioned were obtained after initially filling the missing descriptor values for data sets 8 (9) either by 10 (11) highly intercorrelated topological indices, or by 10 (11) random pseudodescriptors. Modes 1, 4, and 5 do not use the initial descriptor values and therefore should yield identical results for both alternatives. This is the case, as seen in Table 5. Modes 2 and 3 should fall behind modes 1, 4, and 5 in the case of intercorrelated descriptors but not in the case of noncorrelated descriptors. This is exactly what happened, see entries in parentheses in Table 5.

Data Set 10. Karki and Kulkarni fitted by MLR the antibacterial activities of 50 oxazolidinones.[24] Model A in their study contains 6 descriptors selected from a set of 34, $r^2 = 0.732$. Model B contains 3 descriptors selected from the same set, $r^2 = 0.603$. Model C is a 4-descriptor model of $r^2 = 0.651$, where the descriptors were selected from a pool of 10, a subset itself obtained from the 34-descriptor pool by descriptor selection.[24] Therefore for model C also $M = 34$ is most appropriate. Our mode 1/4/5 simulation results for these cases ($it = 250$) are shown in the upper part of Table 6. For all three original models, the distance between r^2 and mode 1 mhr r^2 is >4.5 SD. Thus models A−C are statistically significant. Model C also passed the y-randomization test in the original paper.

Katritzky et al. fitted the same data, enlarged by another ten compounds that had been used as a test set in the earlier study, in three MLR equations containing 7 descriptors each that were selected, by a procedure contained in CODESSA, from two large descriptor pools (739 and 888 descriptors) or from their union (1627 descriptors).[5] Our corresponding random simulation results are shown in the lower part of Table 6 ($it = 25$).

MLR eq 1 in ref 5 contains 7 descriptors selected from the 1627 descriptor pool and has $r^2 = 0.820$. Mode 1 simulation for (60,7,1627) resulted in mhr $r^2 = 0.8188$ with SD 0.0183. The distance between r^2 and mhr r^2 is 0.07 SD (or 0.48 SD from another series of 50 mode 1 experiments, or 0.1 or 0.3 SD from mode 4 or mode 5 experiments), and the original equation therefore does not fit the data significantly better than (on average) the "best" selection of 7 out of 1627 random pseudodescriptors. In fact, 11 of the 25 mode 1 random experiments resulted in highest $r^2 > 0.820$, with maximum 0.8571, and the minimum among all 25 runs was 0.7916. Since the CODESSA procedure excludes collinear descriptors, the effective number of descriptors in the pool may have been a bit lower than 1627. On the other hand, two of the compounds in data set 10 are identical (S10 and S58), so that actually only 59 compounds were treated in ref 5.

In eq 3 in the same study the same activity data were fitted by a MLR model containing 7 descriptors selected from a subset of the previous one containing 888 descriptors, and $r^2 = 0.795$ was obtained. Our mode 1 simulation for (60,7,888) resulted in mhr $r^2 = 0.7772$, SD = 0.0245. Thus the original model's r^2 is nonsignificantly higher (by 0.73 SD) than what is produced by random on average.

In the same work eq 4 fits the data by means of 7 descriptors selected from the complement subset consisting of 739 descriptors ($r^2 = 0.731$). Our mode 1 simulation for (60,7,739) resulted in mhr $r^2 = 0.7635$. Thus r^2 of the original model is even lower than what is obtained on average in random experiments.

Finally in ref 5 the earlier training set / test set partition was reproduced, and the antibacterial activities of the 50 training set compounds were fitted by a MLR model (eq 5, $r^2 = 0.809$) made of 6 descriptors that were selected from the set of 1627. Our mode 1 simulation for (50,6,1627) gave mhr $r^2 = 0.8350$. Thus again the original r^2 is even lower than what was on average obtained by selection among random models.

Taking all results for models from ref 5 together, eqs 1−5 therein cannot be considered statistically significant according to the random number test. Unfortunately, y-randomization was not performed by the original authors, nor could we perform it due to lack of data. Therefore, one should be very cautious in interpreting the descriptors involved in the original models.

It is interesting to note that the equations in ref 5 were subjected to validation procedures in the original work (leave-one-out crossvalidation for all models, leave-one-third-out crossvalidation for eq 1, predictions for a test set for eq 5), without any problems being detected thereby. This suggests that for safety random simulations should always be done.

Data Set 11. Antiplasmodial activities of 16 cinnamic acid derivatives were fitted by MLR by Gupta et al. Seven 3-descriptor equations were established by descriptor selection from a pool of 33 descriptors. The 6 equations having highest r^2 (between 0.757 and 0.706) were rejected for high intercorrelation of descriptors in the model or for low r^2_{cv}. The seventh equation ($r^2 = 0.689$) was considered best, and for this model "*chance correlation <0.01*" and "*better statistical significance >99%*" were claimed on the basis of conventional F values.[34] Bootstrapping and even predictions for a (small) external test set did not reveal any problems with that model. Additionally, a randomization test was done (no details given), and "*chance correlation <0.01 in the randomized biological activity test revealed that the results were not based on chance correlation*".[34] Our mode 1 simulations ($it = 250$) for (16,3,33) resulted in mhr $r^2 = 0.7079$, SD = 0.0808 (see also mode 4 and 5 results, Table 7). Thus, obviously r^2 of the model proposed as best is lower than what results from pure chance on average, and even the highest r^2 model in the original paper does not describe the data significantly better than chance. Obviously, the effect of selection bias was not considered in the original paper, and the extremely overoptimistic judgment obtained thereby was not doubted by the validation procedures performed including an unspecified randomization test.

Data Sets 12−14. The equations proposed by the same group of researchers to describe the binding affinities to PPARα and PPARγ of some benzoylaminobenzoic acids and their transactivation behavior (data sets 12−14)[35] suffer from the same deficiency as that for data set 11. Our simulation results are likewise shown in Table 7.

Equation 1 in ref 35 is a 3-descriptor model ($M = 32$) for PPARγ binding of 16 compounds (data set 12). The reported $r^2 = 0.808$ is higher than mode 1 mhr r^2 for (16,3,32) (0.7051, SD = 0.0852) by 1.2 SD.

Equation 2 is a 3-descriptor model ($M = 32$) for gene transactivation caused by PPARγ binding of 15 compounds (data set 13). The reported $r^2 = 0.750$ is higher than mode 1 mhr r^2 for (15,3,32) (0.7443, SD = 0.0810) by 0.07 SD.

Equation 3 is a 1-descriptor model ($M = 32$) for PPARα binding of 8 compounds (data set 14). The reported $r^2 = 0.738$ is higher than mode 1 mhr r^2 for (8,1,32) (0.5838, SD = 0.1289) by 1.2 SD.

Thus by the random number test none of these equations can be considered statistically significant. Nevertheless in the original paper for all three equations statistical significance was claimed based on conventional F values, and "*chance <0.001*" was claimed based on a "*randomize biological activity data test*" without any details given.

Data Sets 15 and 16. Prabhakar et al. proposed QSAR equations for the antimycobacterial activity of 11 nitro/acetamido alkenols (data set 15), where two descriptors were selected from pools of 68, 96, or 288 descriptors.[36] The $r^2 = 0.748$ given there for eq 3 ($m = 2$, $M = 96$) appears low compared to mode 1 mhr $r^2 = 0.8644$ resulting from our experiments for (11,2,96) (Table 7). Similarly, for 11 chloro/amino alkenols (data set 16) eq 5 ($m = 2$, $M = 96$) has $r^2 = 0.733$, while our mode 1 experiments resulted in mhr $r^2 = 0.8482$. The situation for the other equations given is similar, so that closer examination seems advisable. Interestingly, for all models given some randomization test was done in the original study (no details reported), and in 100 simulations per model "*none of the identified models has shown any chance correlation*".[36]

3. Appropriate and Inappropriate y-Randomized Procedures. The discrepancy between randomization-based significance claims found in the literature (see examples above) and our simulation results caused us some concern. We suspected that inappropriate procedures were applied in the original work in these cases. In fact, in descriptions of y-randomized procedures the second step, building of models for scrambled y data using untouched x data (original descriptors), is often not detailed. In order to learn about the effect of various procedures, we subjected one and the same given random permutation of y data derived from data set 4 ($n = 144$, $m = 10$, $M = 230$) to three procedures, always using descriptors from the original pool.

(1) Target activity values were calculated according to the exact original QSAR equation (original model), i.e., the system was not allowed to adjust to the new situation that arose from y-scrambling (procedure 1).

(2) Target activity values were calculated according to the best model obtainable using the descriptors from the original model, i.e., the system was given the freedom merely to adjust the regression coefficients to the new situation (procedure 2).

(3) Target activity values were calculated according to the "best" model obtained by a new "best" selection of m out of the M original descriptors (procedure 3). This of course is the appropriate procedure if the original model was arrived at by selecting the "best" combination of m descriptors out of M descriptors, and this is what we understand by y-randomization.

For each procedure the resulting (random) r^2 value was recorded. All this was repeated for 25 independent scramblings of the y data. As shown in Table 8, for each single y-permutation the r^2 values arising from the three procedures differ widely, increasing from procedure 1 over procedure 2 to procedure 3. The same, of course, is true for the averages over 25 independent y-scramblings.

While in practice procedure 1 will not be applied (in fact in it no new model is built), use of procedure 2 instead of procedure 3 (i.e., authors' unawareness of selection bias) is a sufficient explanation for the discrepancies. For instance, Guha and Jurs for a $n = 156$, $m = 4$, $M = 65$ case in 100 y-scrambling runs obtained average $r^2 = 0.02$ (range from 0.01 to 0.10) and commented that this is in close accord with the theoretically expected value of r^2 for a model built from random variables.[40] In fact, for $n = 156$ and $m = 4$ the expected r^2 for a random model *without descriptor selection* is 0.026. Our mode 4 experiments for the same (n,m,M) combination ($it = 100$) gave mhr $r^2 = 0.1263$, SD = 0.0244 (values ranging from 0.0718 to 0.1895), wherefrom for y-randomization (mode 2) a mhr r^2 of approximately 0.1 is to be expected. We thank Dr. Guha for informing us that in their scrambling runs a fixed combination of descriptors was used (those of the original model), i.e., selection bias was not accounted for.[41,42]

Table 8. Experimental r^2 Values Resulting from 25 Random Permutations of the y Data from Data Set 4 ($n = 144$, $m = 10$, $M = 230$), Obtained by Three Different Procedures[a]

	r^2		
random perm. no.	procedure 1	procedure 2	procedure 3
1	$1.3 \cdot 10^{-6}$	0.03928	0.27218
2	0.000278	0.04321	0.28787
3	0.016110	0.07762	0.25008
4	0.000655	0.04064	0.29090
5	$2.3 \cdot 10^{-6}$	0.03338	0.18901
6	0.007169	0.08658	0.24169
7	0.003548	0.06098	0.23121
8	0.005697	0.08419	0.34817
9	0.000112	0.03501	0.23710
10	0.000421	0.05125	0.26548
11	0.011629	0.04340	0.33335
12	0.000138	0.07839	0.30173
13	0.035908	0.14101	0.40248
14	0.000140	0.07086	0.30109
15	0.006059	0.04560	0.34036
16	0.000812	0.06508	0.29116
17	0.000121	0.02604	0.35956
18	0.015730	0.14601	0.33755
19	0.001700	0.02935	0.32174
20	0.000959	0.01173	0.22231
21	0.000183	0.04899	0.27195
22	0.003260	0.04598	0.29894
23	0.004392	0.03789	0.31779
24	0.004043	0.09496	0.28999
25	0.001930	0.05744	0.26652
mean	0.004840	0.05980	0.29081
SD	*0.007996*	*0.03258*	*0.04849*

[a] See text. Descriptors from the original pool were used throughout.

4. Approximate Estimation of Mode 4 Mean Highest Random r^2. Simulations involving descriptor selection as described above are time-consuming, particularly for high m and M. Therefore it is highly desirable to be able to simply calculate mhr r^2 for each (n,m,M) case.

There are ${}^{M}C_m$ combinations of m out of M descriptors. For convenience of notation, we denote this number by N.

Background: A prespecified model with m descriptors (no descriptor selection). Under the null hypothesis (no correlation between descriptors and response), for a prespecified model with m descriptors the F statistic

$$F = \frac{r^2}{1 - r^2} \frac{n - m - 1}{m}$$

is F-distributed with $df_1 = m$ and $df_2 = n - m - 1$ degrees of freedom. Therefrom we conclude that the probability that r^2 is less or equal to any $x \in [0,1]$ is

$$P(x) := P(r^2 \le x) = P\left(F \le \frac{x}{1 - x} \frac{n - m - 1}{m}\right)$$

This is the distribution function of random r^2, shown as the black curve in Figure 1 for the example (16,3,53). The expected mean (indicated as the short black line in Figure 1) is given by $m/(n-1)$.[7]

Model with the best combination of m descriptors selected from M descriptors. Let us now consider, instead of a prespecified model, the best of all N models containing m descriptors, which means the model exhibiting maximal r^2, denoted as $r^2_{\max}$. The distribution function for $r^2_{\max}$ is

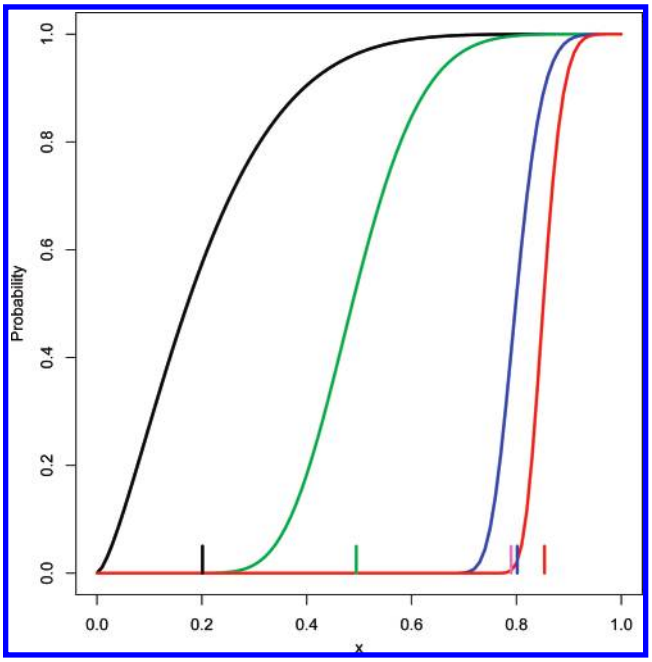

Figure 1. Distribution functions of random r^2 values for the case (16,3,53). Black: models containing three descriptors not selected from a larger pool. Green and red: upper and lower bound curves for models containing three descriptors selected from a pool of 53 (see text). Blue: approximation of the 'true' curve based on eq 5. Short colored lines indicate the expectation of the respective curve. The experimental mhr r^2 (0.7899) is indicated by a pink line.

$$P_m(x) := P(r^2_{\max} \le x) \le P(r^2 \le x) = P(x) \qquad (1)$$

If all models were independent in the sense of not having any descriptor in common (which is true only if $m = 1$ or $m = M$), we had by definition

$$P(r^2_{\max} \le x) = P(x)^N$$

This function is depicted in Figure 1 as the red curve with a short red line indicating its expectation. Unfortunately, this equation does not hold in general, since any two models may contain up to $m - 1$ descriptors in common. Therefore models, instead of being pairwise independent, tend to be positively correlated: The more descriptors are common to two models, the more similar their r^2 values will be. Accordingly, we replace the above equation with the inequality

$$P_m(x) \ge P(x)^N \qquad (2)$$

On the other hand, inequality 1 can be refined by the following consideration. We can choose a subset of $\lfloor M/m \rfloor$ models with disjoint (and thus statistically independent) descriptor sets (where the half square brackets $\lfloor\ \rfloor$ denote the integer part). Thus we obtain

$$P_m(x) \le P(x)^{\lfloor M/m \rfloor} \le P(x) \qquad (3)$$

The function $P(x)^{\lfloor M/m \rfloor}$ is depicted in Figure 1 as the green curve with a green line indicating its expectation. Combining inequations 2 and 3 we conclude that the "true" curve lies between the green curve $P(x)^{\lfloor M/m \rfloor}$ (upper bound) and the red

curve $P(x)^N$ (lower bound):

$$P(x)^N \le P_m(x) \le P(x)^{\lfloor M/m \rfloor} \le P(x) \quad (4)$$

Correspondingly, the "true" mean r^2_{max} lies between the expectations for the green (lower bound) and the red curve (upper bound).

Formula 4 suggests to search for an exponent L with $\lfloor M/m \rfloor \le L \le N$ such that

$$P(x)^L \approx P_m(x)$$

and that $L = 1$ for $m = M$. L is expected to be markedly smaller than N if there is large overlap in the descriptor sets of models.

Rencher and Pun proposed a function $L = (\ln N)^{cN^d}$ and fitted constants c and d for the n,m,M range covered by their experiments.[7]

Salt and Livingstone approximated the 95th percentile of the maximal F value by multiplying the conventional critical F with a so-called inflation index of type N^d, where d is a complex expression containing ten adjustable parameters that were estimated by nonlinear regression.[6b]

Our approach is to define L as a weighted mean of N (assumption of independent models) and 1 (one model only), written as $L = N(1-q) + q$ with a number $q \in [0,1]$. For $q = 0$ (independent models) we obtain $L = N$ (red curve in Figure 1), for $q = 1$ (one model) we obtain $L = 1$ (black curve). Thus, q is a measure of similarity of all N models. We found that q is promising only if it is larger than $(m-1)/m$ and close to 1. The idea is to enlarge $(m-1)/m$ by adding a series of $m-2$ diminishing terms up to a sum close to (but less than) 1 by defining

$$q = \frac{m-1}{m} + \frac{m-2}{m(m-1)} + \frac{m-3}{m(m-1)(m-2)} + \cdots + \frac{1}{m!}$$

This expression can be transformed into $q = 1 - 1/m!$. This provides

$$L = N(1-q) + q = \frac{N}{m!} + \left(1 - \frac{1}{m!}\right) \quad (5)$$

With this choice of L, surprisingly good results are achieved, as shown for the example (16,3,53) by the blue curve in Figure 1. For the n,m,M range covered by our experiments, we found an acceptable agreement between the experimental mode 4 mhr r^2 values and the expectation values of $P(x)^L$ with L defined by eq 5. In Table 9 the experimental mode 4 mhr r^2 values are compared to the expectation values of the black, green, blue, and red curves in Figure 1 type illustrations for various (n,m,M) tuples. The Rencher-Pun approximations are also given.

In Figure 2 the mhr r^2 values estimated by the Rencher-Pun (orange circles) and by our approximation (blue circles) are plotted vs those obtained by computer simulation (mode 4); all data are taken from Table 9. Ideally, all circles should lie on the diagonal (black line).

Note that in contrast to the Rencher-Pun and the Livingstone-Salt approaches our choice of L does not require fitting of several adjustable parameters.

DISCUSSION

For each original model, randomization experiments yield two different numbers for possible comparison with the original r^2, a lower one from randomization experiments that use the original descriptors, such as y-randomization, and a higher one from experiments using random number pseudodescriptors. The difference is caused by the intercorrelation of real descriptors. Intercorrelation of descriptors means that some of them describe similar things, that their vectors point in similar directions, or that m intercorrelated descriptors are equivalent to fewer than m noncorrelated ones.

The question now is which of the two mhr r^2 should be used for a meaningful comparison, or what is meant by the results of both possible comparisons. If the r^2 of an original model significantly exceeds the higher hurdle (situation A), its statistical significance is not in doubt. If, on the other hand, an original r^2 does not exceed the lower hurdle (situation C), the model clearly is not better than random. What, however, if the original r^2 is in between, i.e., if a model passes the y-randomization test but fails in the more demanding random number pseudodescriptor test (situation B)? Shall we rely on the y-randomization result and forget the other? Doing so means to allot decisive importance to the intercorrelation structure of the descriptors in the pool, which also means to esteem these particular descriptors as the only ones that can or should be considered. Knowing that any given pool of descriptors is a more or less arbitrary subset of the set of all possible descriptors available or yet to be developed, we cannot see fundamental importance in the intercorrelation structure within a particular given pool.

We propose the following interpretation of situation B: The descriptors selected for inclusion in the original model obviously catch some major influencing factors, there is a real connection between the selected descriptors and the response, since the latter is not explainable by chance acting on the descriptors in the pool. However, our confidence in the original model is not enhanced by the fact that pure chance, acting without restrictions, produces models that describe the given data as well as or better than the original model and that this happens again and again based on new sets of random number pseudodescriptors. This means that the descriptors in the model and in the pool are insufficient for describing the given property/activity of the given compound set; they lack at least one important dimension.

As demonstrated by the examples in section 2, the phenomenon of selection bias, though known for decades, is still widely ignored. In original reports on MLR modeling authors often are silent on the possibility of chance correlations. Those who do mention such a risk often do not realize the enhanced risk due to descriptor selection, and accordingly by the tests performed models often deceptively seem significant. y-Randomized procedures are sometimes performed incorrectly, i.e., without taking selection bias into account. If selection bias is properly accounted for, that is if descriptor selection is included independently in each y-scrambled run, then we call the procedure y-randomization. It is a useful tool to protect oneself against chance correlation. However, the hurdle built by y-randomization is systematically lower than that built by mode 1 (or mode 4 or 5) simulations as described herein, except in the limiting case of negligible intercorrelation among M descriptors.

y-Randomization requires, along with activity data, knowledge of numerical values of all M descriptors in the pool and therefore, as a rule, is available to the authors of an original model only. In contrast, mode 1/4/5 simulations are

Table 9. Comparison of mhr r^2 for Cases without (left, $M = m$) and with Descriptor Selection (right), for Some (n,m,M) Tuples[c]

			no descriptor selection (n,m,m)	with descriptor selection (n,m,M)				
n	m	M	black curve $m/(n-1)$	green curve	expt (mode 4)[a]	blue curve	red curve	Rencher-Pun approximation
				Data Set 1				
23	3	18	0.136	0.278	0.4323	0.489	0.580	0.416
				Data Set 2				
16	3	53	0.200	0.494	0.7899	0.801	0.853	0.663
				Data Set 3				
48	7	158	0.149	0.311	0.7100	0.706	0.811	0.638
32	7	158	0.226	0.449	0.8965	0.862	0.933	0.807
				Data Set 4				
144	10	230	0.070	0.139	0.3875	0.407	0.519[b]	0.429
129	10	230	0.078	0.154	0.4283	0.444	0.560[b]	0.467
				Data Set 5				
150	14	229	0.094	0.161	0.4638	0.448	0.543[b]	0.593
				Data Set 6				
507	7	249	0.014	0.034	0.0893	0.111	0.143	0.093
				Data Set 7				
82	7	209	0.086	0.195	0.4930	0.520	0.623	0.456
				Data Set 8				
24	4	14	0.174	0.268	0.4251	0.464	0.626	0.466
				Data Set 9				
15	3	14	0.214	0.375	0.5660	0.625	0.732	0.566
				Data Set 10				
50	6	34	0.122	0.204	0.4004	0.429	0.591	0.411
60	7	739	0.119	0.307	0.7638	0.751	0.816[b]	0.701
60	7	888	0.119	0.313	0.7787	0.763	0.816[b]	0.719
60	7	1627	0.119	0.333	0.8119	0.799	0.816[b]	0.776
50	6	1627	0.122	0.368	0.8361	0.821	0.863[b]	0.755
				Data Set 11				
16	3	33	0.200	0.454	0.7191	0.745	0.812	0.624
				Data Set 12				
16	3	32	0.200	0.445	0.7234	0.742	0.809	0.621
				Data Set 13				
15	3	32	0.214	0.472	0.7425	0.770	0.835	0.652
				Data Set 14				
8	1	32	0.143	0.616	0.6117	0.616	0.616	0.426
				Data Set 15				
11	2	96	0.200	0.657	0.8665	0.869	0.890	0.704
				Data Set 16				
11	2	96	0.200	0.657	0.8571	0.869	0.890	0.704

[a] For data sets 1, 2, 8, 9, and 11−16 from Tables 1, 2, 5 and 7; for data sets 3−7 and 10 from additional runs with $it = 100$. [b] This number may suffer from numerical problems due to the high value of the binomial coefficient. [c] Along with mode 4 experimental values the expectations for the black, green, blue, and red curves (Figure 1 type illustrations) are given. The last column shows approximations according to Rencher and Pun (eq 13 in ref 7).

open to everyone if activity data and numbers n, m, and M are known. If activity data also are unavailable, mode 4 simulations are still feasible, yielding the same numerical result as do modes 1 or 5. Mode 4 simulations answer the question how well n random data points would be fitted on average in MLR by the best combination of m out of M random number pseudodescriptors and provide general insight to judge the statistical significance of a MLR model.[6,7]

A factor contributing to the popularity of y-randomization may be the scientists' belief that y-randomization, working on *my response data* (though scrambled) is more relevant for *my problem* than a similar procedure working on random number pseudoresponse data. Comparison between our mode 2 and mode 3 results disproves this belief.

Similarly, one could feel experiments on *my particular response data* to be more relevant than similar experiments on random number pseudoresponse. On the contrary, our experiments showed mode 1 and mode 4 to be numerically equivalent within the limits of random scatter and therefore equally relevant.

In 2004, Estrada et al. wrote "*It is desirable to have as many as possible molecular descriptors to characterize molecular structure but to include as few as possible into the QSAR/QSPR model*" and "*... a larger list of molecular descriptors has more chances of describing better the molecular structure than a shorter one, which means that quantity can be transformed into quality.*"[43] In our opinion, a note of caution has to be amended here: A larger list of molecular descriptors in the pool to select from greatly enhances the danger of chance correlations also.

Program Availability. The random simulations were done using an add-on "RandomQSPR" to be used in connection

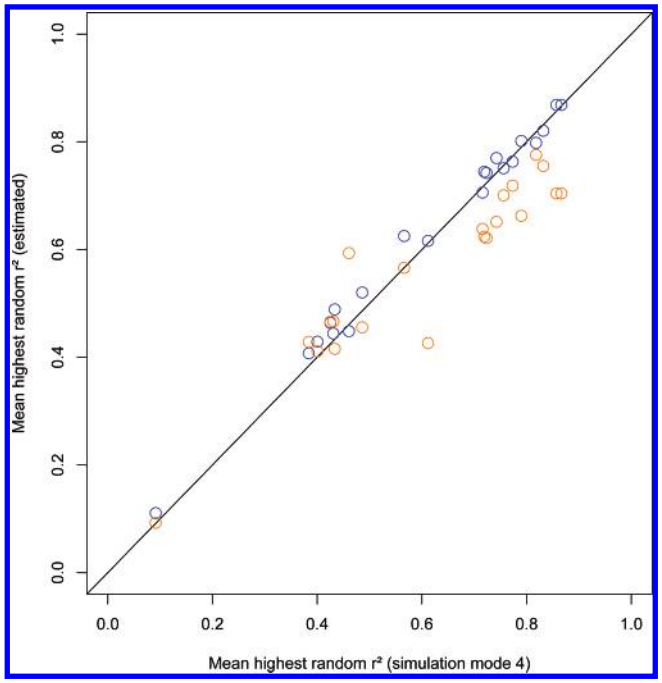

Figure 2. Mean highest random r^2 values estimated by using the Rencher-Pun (orange circles) and our (blue circles) approximation are plotted vs those obtained by computer simulation (mode 4); data are taken from Table 9.

with MOLGEN-QSPR, running on a PC, available from M.M. The theoretical calculations and illustrations (black, green, blue, and red curves and their expectation values as in Figure 1) are obtained using an R program written by and available from G.R. R is a freely available statistics package.[44]

REFERENCES AND NOTES

(1) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066−1068.

(2) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(3) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.-D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of Ambergris Fragrance Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582−595.

(4) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.

(5) Katritzky, A. R.; Fara, D. C.; Karelson, M. QSPR of 3-Aryloxazolidin-2-one Antibacterials. *Bioorg. Med. Chem.* **2004**, *12*, 3027−3035.

(6) (a) Livingstone, D. J.; Salt, D. W. Judging the Significance of Multiple Linear Regression Models. *J. Med. Chem.* **2005**, *48*, 661−663. (b) Salt, D. W.; Ajmani, S.; Crichton, R.; Livingstone, D. J. An Improved Approximation of the Critical *F* Values in Best Subset Regression. *J. Chem. Inf. Model.* **2007**, *47*, 143−149.

(7) Rencher, A. C.; Pun, F. C. Inflation of R^2 in Best Subset Regression. *Technometrics* **1980**, *22*, 49−53.

(8) S. Wold, Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191−193.

(9) Rücker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 2070−2076.

(10) Rücker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points. *J. Chem. Inf. Model.* **2005**, *45*, 74−80.

(11) Rücker, C.; Scarsi, M.; Meringer, M. 2D QSAR of PPARγ Agonist Binding and Transactivation. *Bioorg. Med. Chem.* **2006**, *14*, 5178−5195.

(12) Kubinyi, H. QSAR in Drug Design. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, pp 1532−1554.

(13) Clark, R. D.; Sprous, D. G.; Leonard, J. M. Validating Models Based on Large Data Sets. In *Rational Approaches to Drug Design*, Proceedings of the 13th European Symposion on Quantitative Structure-Activity Relationship, Düsseldorf, Aug 27−Sept 1, 2000; Höltje, H.-D., Sippl, W., Eds.; Prous Science: Barcelona, Spain, 2001; pp 475−485.

(14) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

(15) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(16) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1−12.

(17) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579−586.

(18) Baumann, K.; Stiefl, N. Validation Tools for Variable Subset Regression. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 549−562.

(19) Harrell, F. E. *Regression Modeling Strategies*; Springer: New York, 2001; p 94.

(20) Manly, B. F. J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd ed.; Chapman & Hall: London, 1997; pp 156 and 168.

(21) Miller, A. *Subset Selection in Regression*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, 2002.

(22) Klopman, G.; Kalos, A. N. Causality in Structure-Activity Studies. *J. Comput. Chem.* **1985**, *6*, 492−506.

(23) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; Wiley-VCH: Weinheim, 1995; pp 309−318.

(24) Karki, R. G.; Kulkarni, V. M. Three-Dimensional Quantitative Structure-Activity Relationship (3D-QSAR) of 3-Aryloxazolidin-2-one Antibacterials. *Bioorg. Med. Chem.* **2001**, *9*, 3153−3160.

(25) Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson, L. Model Validation by Permutation Tests: Applications to Variable Selection. *J. Chemometrics* **1996**, *10*, 521−532.

(26) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure-Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using *k* Nearest Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.

(27) Good, P. *Permutation Tests*; Springer: New York, 1994.

(28) Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. A. Consensus kNN QSAR: A Versatile Method for Predicting the Estrogenic Activity of Organic Compounds in Silico. A Comparative Study with Five Estrogen Receptors and a Large, Diverse Set of Ligands. *Environ. Sci. Technol.* **2004**, *38*, 6724−6729.

(29) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.

(30) Kier, L. B.; Hall, L. H. Structure-Activity Studies on Hallucinogenic Amphetamines Using Molecular Connectivity. *J. Med. Chem.* **1977**, *20*, 1631−1636.

(31) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136−142.

(32) Prabhakar, Y. S.; Gupta, M. K.; Roy, N.; Venkateswarlu, Y. A High Dimensional QSAR Study on the Aldose Reductase Inhibitory Activity of Some Flavones: Topological Descriptors in Modeling the Activity. *J. Chem. Inf. Model.* **2006**, *46*, 86−92.

(33) Garg, R.; Kurup, A.; Mekapati, S. B.; Hansch, C. Cyclooxygenase (COX) Inhibitors: A Comparative QSAR Study. *Chem. Rev.* **2003**, *103*, 703−731.

(34) Gupta, A. K.; Soni, L. K.; Hanumantharao, P.; Sambasivarao, S. V.; Arockia Babu, M.; Kaskhedikar, S. G. 3D-QSAR Analysis of Some Cinnamic Acid Derivatives as Antimalarial Agents. *Asian J. Chem.* **2004**, *16*, 67−73.

(35) Hemalatha, R.; Soni, L. K.; Gupta, A. K.; Kaskhedikar, S. G. QSAR Analysis of 5-Substituted 2-Benzoylaminobenzoic Acids as PPAR Modulator. *E. J. Chem.* **2004**, *1*, 243−250. http://cc.1asphost.com/namfarook/newejc/ VOL1/oct04/243.asp (accessed June 20, 2007).

(36) Gupta, M. K.; Sagar, R.; Shaw, A. K.; Prabhakar, Y. S. CP-MLR Directed Studies on the Antimycobacterial Activity of Functionalized Alkenols − Topological Descriptors in Modeling the Activity. *Bioorg. Med. Chem.* **2005**, *13*, 343−351.

(37) Livingstone, D. J.; Salt, D. W. Variable Selection − Spoilt for Choice? *Rev. Comput. Chem.* **2005**, *21*, 287−348.

(38) Stroustrup, B. *The C++ Programming Language*, 3rd ed.; Addison-Wesley: Boston, MA, 2000.

(39) In the original paper it is remarked that *E*(HOMO) and Hammet σ constants were considered as additional descriptors but did not play

any significant role. Our randomization experiments using $M = 24$, i.e. 6 additional descriptors in the pool (1 for E(HOMO) and 5 for σ constants of substituents in 5 ring positions), resulted in somewhat higher mhr r^2 but did not change the overall picture.

(40) Guha, R.; Jurs, P. C. Development of QSAR Models to Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.

(41) E-mail message of Dr. R. Guha to C.R.

(42) The English word "selection" means both the procedure of selecting things and the result of this procedure, i.e. the set of selected things. A phrase such as "using the same descriptor selection as in establishing the original model", referring to the procedure, is therefore easily misunderstood to mean the set of selected descriptors.

(43) Estrada, E.; Delgado, E. J.; Alderete, J. B.; Jana, G. A. Quantum-Connectivity Descriptors in Modeling Solubility of Environmentally Important Organic Compounds. *J. Comput. Chem.* **2004**, *25*, 1787−1796.

(44) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2006. See also http://www.R-project.org (accessed June 20, 2007).

CI700157B