Using Molecular Quantum Similarity Measures under Stochastic Transformation To Describe Physical Properties of Molecular Systems

Xavier Gironés and Ramon Carbó-Dorca*

Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Catalonia, Spain

Received September 26, 2001

The application of molecular quantum similarity measures (MQSM) to correlate physicochemical properties is reported. Satisfactory quantitative structure−property relationship (QSPR) models are obtained for three molecular sets, where boiling points and chromatographic retention times and indices are studied. In this work, MQSM are scaled using a stochastic transformation and related to molecular properties using the partial least-squares technique.

INTRODUCTION

One of the most promising subjects in present day computational chemistry is the characterization and quantification of molecular properties by means of theoretical descriptors. Given a set of molecules and a related molecular property, a quantitative structure−property or −activity relationship (QSPR/QSAR) model can be constructed by deriving descriptors from molecular structure, which will be related to the molecular feature by some statistical technique.

These disciplines were first started by Cross in 1863,[1] when he observed that toxicity of alcohols to mammals increased as water solubility of such alcohols decreased. Since then, QSPR/QSAR techniques have been largely developed using a variety of parameters as molecular descriptors of diverse structural characteristics. For example, Hammett σ values[2] are often used as electronic parameters, and other parameters have been devised to account for the shape, size, lipophilicity, polarizability, and many other molecular structural features. In addition, the recent and fast development of computer architectures has increased the computational power to stages that allow the application of the quantum theory to regular organic molecules at fairly accurate computational levels (semiempirical or ab initio methods with appropriate basis sets) within reasonable time limits and affordable costs. A number of reviews have been published[3−5] concerning the historical development, generation of descriptors, and different methodologies in the QSPR/QSAR fields.

Arising from the conception of a molecule provided by quantum chemistry, and according to the postulate that its electronic density function contains all accessible information contained within this molecule, molecular quantum similarity measures (MQSM)[6−12] stand as a general efficient tool to solve actual chemical problems. MQSM methodology has been successfully applied within pharmacological[13−18] and toxicological[19−21] problems. MQSM are based on the concept of molecular similarity[22,23] and are related to a self-evident molecular similarity principle: "the more similar two molecules are, the more similar properties they will possess". This last statement requires a procedure to compare the molecules, and MQSM easily establish the degree of similarity on the basis of the electronic distribution, namely first-order density functions (DF), of the molecular structures.

In this work, it is intended to use the MQSM methodology to correlate boiling points and chromatographic parameters of simple organic molecules. For this purpose, this paper is structured as follows: a survey around the employed protocol is given first, next the presentation of the molecular sets, as well as the results achieved, are described, and the concluding remarks are finally given.

MATERIAL AND METHODS

MQSM and Stochastic Transformation. The practical application of MQSM relies on the use of first-order density functions. These functions are quantum-mechanical observable elements producing information on the molecular electron distribution. Within this framework, two molecular structures are considered to be similar if their electron distributions are similar. Thus, among other possibilities,[8] quantitative measures of the similarity between two molecules can be defined as the direct volume integral between their density functions weighted by the *Coulomb* operator:

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)|\mathbf{r}_1 - \mathbf{r}_2|^{-1}\rho_B(\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2 \qquad (1)$$

Here $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ are the first-order electron density functions of molecules A and B, respectively, and Z_{AB} is the resulting quantum similarity measure. As MQSM depend on the relative position of both compared objects in space, a molecular alignment procedure is required. When studied molecular sets share common structural features, the topogeometrical superposition algorithm (TGSA)[24] is used, as it performs pairwise superpositions according to the molecular backbones. Once superposed, the overall set of MQSM for a series of N molecules is computed and collected in the way of an ($N \times N$) similarity matrix form, which can be

* Corresponding author phone: (+34)972418367; fax: (+34)972418356; e-mail: quantum@iqc.udg.es.

10.1021/ci0103370 CCC: $22.00
Published on Web 03/06/2002

used to extract the appropriate correlation information and build QSPR/QSAR descriptors. Once the similarity matrix has been computed, a possible scaling of the MQSM can be done by means of eq 2, providing the recently described s*tochastic transformation*[25] of MQSM. The construction of this stochastic MQSM matrix conceives that the sum of every row elements can be used as scale factors to obtain a new row. This procedure originates a new nonsymmetric square matrix, which will act as a set of input descriptors in any subsequent statistical procedure.

$$S_{ij} = Z_{ij}\left(\sum_{j}^{N} Z_{ij}\right)^{-1} \tag{2}$$

Promolecular Atomic Shell Approximation (ASA). To avoid expensive theoretical calculations, the promolecular atomic shell approximation (ASA)[26−28] has been used here to construct molecular first-order electron density. Within ASA the molecular density is expressed as a sum of discrete atomic density contributions, which are taken as a superposition of 1*S* Gaussian functions and fitted to atomic ab initio ones. Once the overall atomic densities are built, each molecular density function can be constructed by adding as appropriate these elementary atomic building blocks. Since it has been proved that the MQSM from fitted densities built in this way differ by up to a 2% from the ab initio ones,[27] their use is clearly justified.

Molecular Alignment. MQSM are dependent on the relative position of the molecules under comparison. In this way, the topo-geometrical approach (TGSA)[24] has been used to perform the needed pairwise molecular alignments. This molecular superposition method overlays the involved molecules according to the maximal common substructure shared by the analyzed compounds.

Molecular Modeling. All molecules have been constructed using the Ampac 6.55[29] software package. The geometry optimization of the resulting structures has been carried out in gas phase with the same program at the AM1[30] computational level.

Treatment of Quantum Similarity Matrixes and Model Building. Common chemometric tools may be applied to deal with similarity matrices. Particularly, partial least squares (PLS)[31,32] stands as an ideal technique to obtain a generalized regression to model the association between the matrices **X** (descriptors) and **Y** (responses). In computational chemistry, its main use is to model the relationship between computed variables, which together characterize the structural variation of a set of *N* compounds and a property of interest measured on those *N* substances.[33−35] This variation of the molecular skeleton is condensed into the matrix **X**, whereas the analyzed properties are recorded into **Y**. In PLS, the matrix **X** is commonly built up from nonindependent data, as it uses more columns than rows; hence it is not called the *independent* matrix but a predictor or descriptor matrix. A good review, as well as its practical application in QSAR, is found in ref 36, and a detailed tutorial in ref 37.

Unlike regression, PLS is not based on the assumption of independent and precise **X** variables, but it is rather based on the more realistic assumption that **X** contains more or less collinear and noisy parameters. PLS summarizes these **X** variables by means of a few orthogonal score vectors ($\mathbf{t}_a \in \mathbf{T}$), and the matrix **Y** is also resumed in few score vectors ($\mathbf{u}_a \in \mathbf{U}$) that are not orthogonal. Plots of columns from **T** and **U** provide visual representation of the configuration of the observations in the **X** or **Y** spaces, respectively. To speed up the PLS routine, the dimensionality of the similarity matrix is a priori reduced in two ways:

(a) removal of columns with a variation coefficient below 2%, thus eliminating those descriptors that are roughly a vector of constants;

(b) extraction of all columns that, after pairwise corrrelation, score a goodness-on-fit coefficient larger than 0.85, thus eliminating redundant descriptors.

The PLS procedure allows one to derive a number of *factors* and *weights*, which are used to describe the desired properties. QSPR models are built up from these factors and weights.

In this work, all obtained models are evaluated by commonly used statistical parameters: goodness-on fit (r^2);[38] root-mean-square error between experimental and predicted values (s)[38] by *leave-one-out*;[39] predictive capacity (q^2).[38] In addition, and in order to avoid chance correlations and excess of parameters, models are submitted to random tests, where the properties are randomly permuted in their positions and the entire modeling procedure is repeated a number of times. If satisfactory correlations are found within the random test, the model obtained should not be trusted, as the methodology used may be potentially capable to correlate any kind of data.

RESULTS AND DISCUSSION

In this section, it is intended to prove the usefulness of the proposed procedure with three different molecular sets. The molecular properties analyzed consist of boiling points, gas chromatographic retention times, and gas−liquid chromatographic retention indices, respectively. The boiling points are analyzed for a molecular set composed of 529 saturated hydrocarbons[40] (linear, branched, and cyclic). The retention times are contemplated for a set of 152 diverse chemical compounds.[41] Finally, the retention indices are evaluated over a set of 50 phenol derivatives.[42]

Boiling Point for 529 Saturated Hydrocarbons. Boiling point at normal pressure (bp) has become a benchmark property for the evaluation of novel QSPR descriptors.[43,44] In this way, it is interesting to test the present proposed MQSM methodology for such a homologous molecular series. For this purpose, a set of 529 saturated hydrocarbons, whose data are presented in Table 1, was tested for correlation search with their bp using the previously described methodology. The initial 529 × 529 matrix was reduced to 529 × 513 after the redundant variable reduction. QSPR models have been constructed up to the usage of 15 PLS factors, in which both r^2 and q^2 have been inspected. The optimal number of parameters to be used in the final model, which in this case has been fixed to 6 PLS factors, has been chosen from the evolution of these magnitudes, as presented in Figure 1.

As observed in Figure 1, a sharp increase in the predictive capacity occurs up to the sixth PLS factor and a slow increase until the eighth one. A further increase in the number of parameters results, in despite of the expected increase of r^2, in a decrease of the predictive power of the generated model. In this way, the chosen number of PLS factors used was 6,

Table 1. SMILES Notation[47,48] and Boiling Points (°C) of 529 Saturated Hydrocarbons[40]

compd	bp	compd	bp	compd	bp
CCC	32.8	CCCCCCCCC	99.2	CCCC(C)(C)CCCC	172.5
C_1CC_1	42.1	CC(C)CCCCCC	118.2	CCC(CC)C(CC)CC	166
CCCC	8	CCC(C)CCCCC	109.8	CC(C)(C)CC(C)CCC	164.5
CC(C)C	0.7	CCCC(C)CCCC	114.8	CC(C)C(C)C(C)CCC	179.5
$C_1(C)CC_1$	12.6	CC(C)CCCC(C)C	106.5	CC(C)CCC(C)(C)CC	171.5
C_1CCC_1	0.5	CCCCC(CC)CC	140.5	CC(C)C(CCC)C(C)C	175.8
$C_{12}CC_1C_2$	11.7	CCCC(CC)CCC	168.3	CCC(C)C(C)C(C)CC	160.5
CCCCC	36	CC(C)CCC(C)CC	164.6	CC(C)(C)CC(CC)CC	165
CC(C)CC	46	CC(C)C(C)CCCC	162.5	CCC(C)(CC)CCCC	172.5
CC(C)(C)C	39	CCC(C)C(C)CCC	166	CC(C)C(C)C(CC)CC	186.2
$C_1(CC_1)CC$	33.5	CCC(C)CC(C)CC	161.4	CC(C)C(CC)C(C)CC	174.5
$C_1(C)CC_1C$	35.9	CCC(CC)CC(C)C	142.5	CC(C)CC(C)(C)CCC	164
$C_1(C)(C)CC_1$	32.6	CC(C)(C)CCCCC	147.8	CC(C)(C)C(C)CCCC	160.5
$C_1(C)CCC_1$	20.6	CCC(C)C(C)CCC	140.7	CCC(C)(C)CC(C)CC	159.5
C_1CCCC_1	36.3	CCCC(CC)C(C)C	161	CCCC(C)(CC)CCC	162
$C_1C_2CC_1C_2$	49.3	CC(C)CC(C)CCC	146	CC(C)(C)CC(C)C(C)C	172.5
$C_1CC_2CC_{12}$	36	CCC(C)CC(C)CC	137.5	CC(C)(C)CCC(C)(C)C	165.5
$C_1CC_{12}CC_2$	27.8	CC(C)(C)CCC(C)C	137.5	CC(C)C(C)C(C)C(C)C	159.5
$C_{12}(C)CC_1C_2$	9.5	CCC(C)(C)CCCC	131	CC(C)C(C)(C)CCCC	163
CCCCCC	71	CCCC(C)(C)CCC	128	CC(C)CC(C)(CC)CC	150
CC(C)CCC	76	CC(C)C(C)C(C)CC	152.5	CC(C)(C)C(CC)CCC	168
CCC(C)CC	83	CC(C)(C)CC(C)CC	169	CC(C)(C)C(C)CC(C)C	153
CC(C)C(C)C	81	CC(C)C(CC)C(C)C	161.8	CCC(CC)(CC)CCC	168.5
CC(C)(C)CC	69.5	CCC(C)(CC)CCC	163	CC(C)C(C(C)C)C(C)C	160
$C_1(CC_1)CCC$	60.5	CC(C)CC(C)(C)CC	163	CCC(C)(C)C(C)CCC	152
$C_1(CC_1)C(C)C$	55	CC(C)(C)C(C)CCC	164	CCC(C)C(C)(C)CCC	157.5
$C_1(CC)CC_1C$	69	CCC(CC)(CC)CC	157.8	CC(C)C(C)(C)CC(C)C	158.5
$C_1(C)(CC_1)CC$	58.3	CC(C)C(C)(C)CCC	168	CCC(C)(C)C(CC)CC	159.5
$C_1(C)C(C)C_1C$	63	CC(C)(C)C(CC)CC	155	CC(C)C(C)(CC)CCC	152
$C_1(C)(C)CC_1C$	57	CCC(C)(C)C(C)C	154	CC(C)(C)CC(C)(C)CC	155.1
$C_1(CC)CCC_1$	63	CC(C)(C)C(C)(CC)CC	152.5	CC(C)(C)C(C)C(C)CC	148.3
$C_1(C)CC(C_1)C$	52.6	CC(C)(C)CC(C)(C)C	157	CCC(C)(CC)C(C)CC	153.5
$C_1(C)CCC_1C$	70.7	CC(C)(C)C(C)C(C)C	159	CC(C)(C)C(CC)C(C)C	146.7
$C_1(C)(C)CCC_1$	59	CC(C)C(C)(C)C(C)C	149	CC(C)C(C)C(C)(C)CC	146
$C_1CCC(C_1)C$	62	CC(C)(C)C(C)(C)CC	153	CC(C)C(CC)(CC)CC	158
C_1CCCCC_1	53.6	$C_1(CC_1)CCCCCC$	145	CC(C)C(C)(C)C(C)CC	155.5
$C_{12}CCC(C_1)C_2$	71.8	$C_1(CC_1)C(C)CCCC$	152	CC(C)C(C)(CC)C(C)C	154
$C_1CC_1C_2CC_2$	80.7	$C_1(CC_1C)CCCCC$	148.2	CC(C)(C)C(C)(C)CC	145.5
$C_1CC_2C_1CC_2$	68.7	$C_1(CC)CC_1CCC$	151	CC(C)(C)C(C)C(C)(C) C	180
$C_{12}CCCC_1C_2$	60.3	$C_1CC_1(CC)CCCC$	146.7	CCC(C)(C)C(C)(C)CC	172
$C_1CCC_{12}CC_2$	63.3	$CC(C)(C)CC_1(C)CC_1$	151.5	CC(C)(C)C(C)(CC)CC	170.7
$C_{12}(C)CCC_1C_2$	58	$CC(C)CC_1CC_1(C)C$	149	CC(C)(C)C(C)(C)C(C) C	176.5
$C_{12}(C)CC_1(C_2)C$	49.7	$C_1(C)(C)CC_1C(C)(C)C$	151	$C_1(C)(CC_1)CCCC(C)C$	172
CCCCCCC	108.5	$C_1(C)(CC)CC_1(C)CC$	143	$CC(C)CC_1CC_1C(C)C$	153
CC(C)CCCC	104	$C_1(C)(C)C(C)(C)C_1(C)C$	138.8	$C(C)CCC_1CC_1(C)C$	175.7
CCC(C)CCC	105	$C_1CCC_1C(CC)CC$	146.5	$CCCC_1CC_1(C)CCC$	173.9
CCC(CC)CC	107	C_1CCCC_1CCCC	142	$CC_1(C)CC_1C(C)(C)CC$	173.5
CC(C)CC(C)C	110	$C_1CCCC_1CC(C)C$	148.2	$C_1(C(C)C_1(C)C)C(C)(C)C$	160
CC(C)C(C)CC	107.5	$C_1C(C)CCC_1CCC$	150	$C_1(CCC_1C(C)C)C(C)C$	172
CC(C)(C)CCC	105.5	$C_1CCCC_1C(C)CC$	150.5	$C_1(CC)C(C)C(C)C_1CC$	173.9
CCC(C)(C)CC	110	$C_1C(CC)CCC_1CC$	132.2	$C_1(C)(C)C(CC)CC_1CC$	160.5
CC(C)(C)C(C)C	110.5	$C_1(C)CCCC_1CCC$	154.5	$C_1(C)C(CC_1(C)C)C(C)C$	148.5
C_1CC_1CCCC	116	$C_1(CC)CCCC_1CC$	141	$C_1CCC(C_1)CCCCC$	172
$C_1CC_1C(C)CC$	96.5	$C_1C(C)CCC_1C(C)C$	131	$C_1CCC(C_1)CCC(C)C$	171
$C_1(C)CC_1CCC$	98.5	$C_1(C)CCCC_1C(C)C$	135	$C_1(C)CCC(C_1)CCCC$	163
$C_1(CC)CC_1CC$	100	$C_1CCCC_1(C)CCC$	127.2	$C_1CCC(C_1)C(C)CCC$	164
$C_1(C)(CC_1)CCC$	103	$C_1C(C)C(C)CC_1CC$	149	$C_1CCC(C_1C)CCCC$	141.5
$C_1(CC_1C)C(C)C$	81.5	$C_1C(C)CC(C)C_1CC$	143	$C_1CC(CC_1C)CC(C)C$	180.9
$C_1(CC_1)C(C)(C)C$	92	$C_1CC(C)C(C)C_1CC$	153	$C_1CCCC_1C(CC)CC$	171.3
$C_1(CC_1)(CC)CC$	71.5	$C_1CC(C)C(C_1C)CC$	140.2	$C_1CCCC_1CC(C)CC$	173.4
$C_1(CC)C(C)C_1C$	74	$C_1(CCCC_1)C(C)(C)C$	125	$C_1CCCC_1C(C)C(C)C$	169
$C_1(C)(CC_1)C(C)C$	78	$C_1(CC)(CC)CCCC_1$	125.5	$CC(C)C_1CCCC_1CC$	179.3
$C_1(C)(C)CC_1CC$	84	$C_1(CC)CCC(C_1)(C)C$	121	$C_1(CC)CCC(C_1C)CC$	175.5
$C_1(C)(CC)CC_1C$	98	$C_1(C)CCC(C_1)(C)CC$	130.8	$CC(C)(CC)C_1CCCC_1$	172
$C_1(C)(C)C(C)C_1C$	90.3	$C_1CCC(C_1)(C)C(C)C$	124.5	$CC(C)C_1CC(C)C(C)C_1$	170
$C_1(C)(C)CC_1(C)C$	93	$C_1CC(CC)C(C_1)(C)C$	148.7	$CC(C)C_1CCC(C)(C)C_1$	176
C_1CCC_1CCC	90	$C_1C(C)C(C)C(C)C_1C$	156.6	$C_1(C)C(C)CC(C_1C)CC$	167
$C_1CCC_1C(C)C$	84.9	$C_1CCC(C)C_1(C)CC$	148	$C_1(CC)C(C)CC(C)C_1C$	168.5
$C_1C(C)CC_1CC$	81.1	$C_1C(C)C(C)CC_1(C)C$	148.3	$C_1(CCCC_1(C)C)C(C)C$	171
$C_1CC(C)C_1CC$	80.5	$C_1(C)CC(C)CC_1(C)C$	154.3	$C_1(C)C(C)C(C)C(C)C_1C$	174.3
C_1CCCC_1CC	88.6	$C_1(C)C(C)CCC_1(C)C$	148.2	$C_1C(C)(C)CC(C)C_1(C)C$	175
$C_1C(C)CCC_1C$	91	$C_1C(C)(C)CCC_1(C)C$	149.5	$C_1CCCC(C_1)CCCC$	171.5
$C_1(C)CCCC_1C$	81.5	$C_1(C)CCC(C)C_1(C)C$	150.5	$C_1CCCC(C_1)CC(C)C$	179.5
$C_1CCCC_1(C)C$	79.1	$C_1(C)(C)CCCC_1(C)C$	141	$C_1CC(CCC)CCC_1C$	168

Table 1. (Continued)

compd	bp	compd	bp	compd	bp
$C_1CCCC(C_1)C$	85.2	C_1CCCCC_1CCC	145	$C_1(C)CCCC(C_1)CCC$	166.6
$C_1CCCCCC_1$	78	$C_1CCCCC_1C(C)C$	145	$C_1CCCC(C_1)C(C)CC$	166.5
$C_1CC_1CCC_2CC_2$	76	$C_1CC(C)CCC_1CC$	142.5	$C_1CC(CC)CCC_1CC$	177.5
$C_1CC_2CCC_1C_2$	100.7	$C_1C(C)CCCC_1CC$	138	$C_1CC(CC)CC(C_1)CC$	172.5
$C_1CC_2CC(C_1)C_2$	92.7	$C_1(C)CCCCC_1CC$	147	$C_1(CC)CCCCC_1C$	153
$C_{12}CCCC_1CC_2$	89.5	$C_1(C)CC(C)CC(C_1)C$	151	$C_1CC(CCC_1C)C(C)C$	160.3
$C_{12}CCCCC_1C_2$	94	$C_1(C)CCC(C)C(C_1)C$	145	$C_1CCCC(CC)C_1CC$	158
$C_1CCC_{12}CCC_2$	103.5	$C_1CCC(C)C(C)C_1C$	151	$C_1C(C)CCCC_1C(C)C$	153
$C_1CCCC_{12}CC_2$	91.3	$C_1CCC(C)(CC_1)CC$	133	$C_1C(C)CC(C)CC_1CC$	155
$C_{12}CCC(C)C_1C_2$	95.6	$C_1(C)CCC(C)(C)CC_1$	135.5	$C_1(C)CCCCC_1C(C)C$	167
$C_{12}CCCC_1C_2C$	97.9	$C_1CCC(C)(C)CC_1C$	143	$C_1CCCCC_1(C)CCC$	161.5
$C_{12}CCC(C_1)(C_2)C$	101	$C_1CCC(C)(C)C(C_1)C$	138	$C_1(C)C(C)CCCC_1CC$	182.8
$C_{12}CCCC_1(C_2)C$	118.4	$C_1CCCCC(C_1)CC$	135.9	$C_1(CCCCC_1)C(C)(C)C$	164.2
$C_1C_2(C)CC_1(C)C_2$	98.5	$C_1CC(C)CCC(C_1)C$	142.5	$C_1(CC)(CC)CCCCC_1$	173
$C_{12}(C)CCC_1(C_2)C$	90	$C_1CCC(C)CC(C_1)C$	127	$C_1CC(C)CCC_1(C)CC$	170.7
$C_1CC_{12}CC_2(C)C$	92	$C_1CCCC(C)C(C_1)C$	129.5	$C_1C(C)C(C)CC(C)C_1(C)C$	177.5
$C_{12}(C)CC_1C_2(C)C$	93.5	$C_1CCCCC(C_1)(C)C$	134	$C_1C(C)CCCC_1(C)CC$	161
$C_{12}CC_2CC_3C_1C_3$	80.5	$C_1CCCC(C)CCC_1$	118.2	$C_1(C)C(C)CC(C)CC_1C$	174
$C_{12}CC_1C_3CC_2C_3$	89.8	$C_1CCCCCCCC_1$	138	$C_1CCCCC_1(C)C(C)C$	191.4
$C_{12}CCC_3C(C_1)C_3C_2$	79.2	$C_1CCC_2CC_1CCC_2$	135	$C_1(C)C(C)CCC(C)C_1C$	183.5
$C_{12}CCCC_3C_1C_{23}$	86.1	$C_1CCC_1CC_2CCC_2$	156.7	$C_1(C)(C)CC(C)CC(C_1)C$	181.5
$C_{123}CC_1CCC_2C_3$	80.9	$C_1CCCC_2CC_1CC_2$	154.8	$C_1(C)(C)CCC(C)C(C_1)C$	185.5
$C_{12}CC_3C_4C_1C_2C_{34}$	137.5	$C_1CCCC_2CCCC_{12}$	150.8	$C_1(C)CCC(C)CC_1(C)C$	170.5
$C_{12}CC_1C_3C_4C_2C_{34}$	153	$C_1CCCC_2CCC_2C_1$	150	$C_1(C)(C)CCC(C)(C)CC_1$	193.6
CCCCCCCC	106	$C_1(CCCCC_1)C_2CC_2$	154.3	$C_1(C)(C)CCCC(C_1)(C)C$	142
CC(C)CCCCC	142	$C_1CCCCC_2CC_2C_1$	139.5	$C_1(C)(C)CCCC(C)C_1C$	174.1
CCC(C)CCCC	149	$C_1C(C)C_2CCC_1C_2$	144.8	$C_1(C)(C)CCCCC_1(C)C$	166.8
CCCC(C)CCC	136	$C_1CC_2CCC(C_1C)C_2$	149.4	$C_1CCCCC(C_1)CCC$	167.8
C(C)(C)CCC(C)C	125	$C_1CC_2CC(C_1)C(C)C_2$	152	$C_1C(C)CCC(C)CC_1C$	165.7
C(C)CC(CC)CC	120.5	$C_1C(C)CC_2C_1CCC_2$	136	$C_1CC(C)C(C)CCC_1C$	160
C(C)(C)CC(C)CC	103	$C_{123}CCCC_1{\cdot}C_2CCC_3$	136.6	$C_1CC(C)CC(C)C(C_1)C$	165.1
C(C)(C)C(C)CCC	111	$C_{123}CCC_1{\cdot}C_2CCCC_3$	145.1	$C_1CCC(C)C(C)C(C_1)C$	168
C(C)C(C)C(C)CC	115	$C_1CCC_2CCC(C)C_{12}$	163.7	$C_1CCC(C)(C)CC(C_1)C$	158.5
CC(C)(C)CCCC	102	$C_{123}CC_1{\cdot}C_2CCCCC_3$	153.8	$C_1CCC(C)C(C)(C)CC_1$	164
CCC(CC)C(C)C	129	$C_1CC_2(C)CCC_1CC_2$	151	$C_1CCCCCC(C_1)CC$	157
CC(C)C(C)C(C)C	137	$C_1CC_2CC_1CC_2CC$	157	$C_1CC(C)CCCCC_1C$	161.8
CCC(C)(C)CCC	136	$C_{12}CCCC(C_2)(C)CC_1$	150	$C_1(C)CCCCCC(C_1)C$	153
CC(C)(C)CC(C)C	133	$C_1CC_2CC_1C(C)C_2C$	168.2	$C_1CCCCCC(C)C_1C$	159.5
CCC(C)(CC)CC	141	$C_1CC_2(CC)CC_1CC_2$	175	$C_1CCCCCC(C_1)(C)C$	164
CC(C)(C)C(C)CC	125	$C_1CC_2CC(C_2)C_1(C)C$	157.8	$C_1CCCCC(C)CCC_1$	159.7
CC(C)C(C)(C)CC	128	$C_1CCCC(C)C_{12}CC_2$	174.5	$C_1CCCCCCC_1$	159
C(C)(C)(C)C(C)(C)C	128	$C_1CCCC_2(C)CC_2C_1$	150.8	$C_{12}CCCCC_2CCCC_1$	155
C(CCCC)C$_1$CC$_1$	130.5	$C_1C_2CCC(C_2)C_1(C)C$	142.8	$C_1CCC(C_1)C_2CCCC_2$	160
CC(CCC)C$_1$CC$_1$	125	$C_1C(C)C_2(C)CC_1CC_2$	144	$C_1CCCCC_2CCCC_{12}$	145
C$_1$(CC$_1$C)CCCC	117	$C_1CCCCC_{12}CC_2C$	142.4	$C_{12}CCCC(C_2)CC(C_1)C$	166
C$_1$CC$_1$CC(C)(C)C	138	$C_1CC_2(C)CCC_1C_2C$	134	$C_{12}CCCC(CC_2)C(C_1)C$	160
C$_1$CC$_1$C(C)C(C)C	125	$C_1CC_2CCC_1C_2(C)C$	143	$C_{12}CCCC(C_1)CCC_2C$	155.7
C$_1$(CCC)CC$_1$CC	115	$C_1(C)(C)C_2CCCC_1C_2$	142.1	$C_1C(C)CCC_2C_1CCC_2$	166
CC$_1$CC$_1$CC(C)C	91	$C_1CCCC_2CC_{12}CC$	136	$C_1CCCC_2C_1CC(C)C_2$	162.4
C$_1$(C)(C)CC$_1$CCC	126.1	$C_1(C)CCC_{12}CCC_2C$	133.5	$C_1CCC_2CCCC_1C_2C$	159.5
C$_1$(C)(CC)CC$_1$CC	104	$C_1CCCC_2C_1C_2(C)C$	140.5	$C_1CCCC_2(C_1)CCCC_2$	148.2
C$_1$(C)(C)CC$_1$C(C)C	105	$C_1(C)CCC_2C_1C_2(C)C$	136	$C_1(C)CCCC_2C_1CCC_2$	167
C$_1$(C)(C)CC$_1$(C)CC	159	$C_{12}(C)CCC(C_2)C_1(C)C$	133.8	$C_1CCCCC_2(C_1)CCC_2$	159.7
C$_1$(C)(C)C(C)C$_1$(C)C	128	$C_1CC_2(C)CC_2(C_1)CC$	132.7	$C_{12}CCC(CC_2)CC_1CC$	167
C$_1$(CCC$_1$)CC(C)C	117.7	$C_1(C)CC_2(C)CC_2(C_1)C$	140.6	$C_1CCCC_2CCC(C)C_{12}$	161.2
C$_1$(CCC)CC(C$_1$)C	124	$C_{12}CC_3CC(C_1)C(C_2)C_3$	138	$C_{12}(C)CCCC(CC_1)CC_2$	157
C$_1$(CCC$_1$)C(C)CC	106	$C_1C_2CC_3CCC_1C_3C_2$	131.3	$C_1CCC_2CCC(CC)C_{12}$	147
C$_1$(CC)CCC$_1$CC	115.5	$C_{12}CCC(C_2)CC_3C_1C_3$	140.4	$C_{12}CCC(C_2)CC(C)C_1C$	157
C$_1$(C)C(C)C(C)C$_1$C	108	$C_1CC_2C_3C_1CC_2CC_3$	124	$C_{12}CCC(CC_2C)CC_1C$	160
C$_1$(C)(C)CC(C$_1$)(C)C	110	$C_1CC_2(CC_2)CC_{13}CC_3$	137.3	$CC_1CC_2CC(C)CC_2C_1$	162
C$_1$CCCC$_1$CCC	105.9	$C_1(CC_1)(C_2CC_2)C_3CC_3$	135.2	$C_{12}CC(C)C(C_2)CCC_1C$	147.7
C$_1$CCCC$_1$C(C)C	108.9	$C_1CC_2CCC_1C_{23}CC_3$	139	$C_{12}(CC)CCC(CC_1)CC_2$	163
C$_1$C(C)CCC$_1$CC	94.4	$C_{12}C(C)C_2C_3CC_1CC_3$	126.5	$C_{12}CCC(CC_2)C(C)C_1C$	152.8
C$_1$(C)CCCC$_1$CC	104.5	$C_1CCC_2(CC_2)C_{13}CC_3$	136.7	$C_1CCCC_2CCCC_{12}C$	163
C$_1$(C)CC(C)CC$_1$C	100.5	$C_1C_2CC_3C_1C_3(CC)C_2$	140.6	$C_1CC(C)C_2CCC(C)C_{12}$	164
C$_1$CCCC$_1$(C)CC	120.1	$C_1C_2C_3CC_1C(C)(C)C_{23}$	130.7	$C_1CCC_2C_1C(C)CC_2C$	147
C$_1$(C)C(C)CCC$_1$C	117.4	$C_1C_2CC_3C_1C_3(C)C_2C$	133.6	$C_{12}CCC(C_2)CCC_1(C)C$	163.8
C$_1$C(C)CCC$_1$(C)C	123	$C_1C_2CC_3C_1(C)C_3(C)C_2$	145	$C_{123}CCCC_1{\cdot}C_3CCC_2C$	164
C$_1$(C)CCCC$_1$(C)C	119	$C_{12}CC(C_3C_1C_3)C_4C_2C_4$	137.7	$C_{12}CCC(CC_2)CC_1(C)C$	164
C$_1$CCCC(C$_1$)CC	114.5	$C_{12}CCC_3C_2C_4C_1C_{34}$	133.8	$C_{12}(C)CCC(C)C(C_2)CC_1$	152
C$_1$CC(C)CCC$_1$C	86	CCCCCCCCCC	122.3	$C_1CCC_2CC(C)CC_{12}C$	158
C$_1$CC(C)CC(C$_1$)C	131	CC(C)CCCCCCC	133	$C_{12}(C)CCCC(C_2)(C)CC_1$	155.7
C$_1$CCC(C)C(C$_1$)C	126.4	CCC(C)CCCCCC	141.5	$C_{12}CC(C)C(C_2)CC_1(C)C$	166
C$_1$CCCC(C$_1$)(C)C	121	CCCC(C)CCCCC	140.2	$C_1CCCC_2CC_{12}CCC$	147.9

Table 1. (Continued)

compd	bp	compd	bp	compd	bp
$C_1CCCC(C)CC_1$	124.7	CC(C)CCCCC(C)C	160	$C_{12}CCC(C_2)C(C)C_1(C)C$	137
$C_1CCCCCCC_1$	115	CCCCC(C)CCCC	174	$C_{12}(C)CCC(C_2)C(C)C_1C$	158
$C_1(CC_1)CCC_2CC_2$	121.5	CCC(CC)CCCCC	155	$C_{12}CCC(CC_1C)C_2(C)C$	160
$C_{12}CCCC_1CCC_2$	117	CC(C)CCCC(C)CC	175.5	$C_{12}(C)CCC(C_2)C(C_1)(C)C$	158
$C_1(CCC_1)C_2CCC_2$	104.9	CCCC(CC)CCCC	219	$C_{12}CC(CCC_1C)C_2(C)C$	159
$C_1CCC_2CCC_2C_1$	114	CC(C)CCC(C)CCC	189.5	$C_{12}(C)CCC(C)(CC_1)C_2C$	148.7
$C_1CCCCC_2C_1C_2$	131.8	CCCC(CCC)CCC	158.3	$C_1C(C)CCC_2C_1C_2(C)C$	166.3
$C_1CC_2CC_1C(C)C_2$	121.8	CC(C)CC(C)CCCC	192.7	$C_{12}(C)CCC(CC_1)C_2(C)C$	157
$C_1CCCC_{12}CCC_2$	122.3	CCC(C)CCC(C)CC	160	$C_1CCCCC_{12}CC_2(C)C$	164
$C_{12}CCC(CC_1)C_2C$	126.6	CC(C)C(C)CCCCC	151	$C_1CC(C)C_2CC_{12}C(C)C$	164
$C_1(C)CCC_2C_1CC_2$	119.5	CCC(CC)CCC(C)C	143.5	$C_1CCC_2(C)CC_{12}CCC$	153
$C_{123}CC_1C_2CCCC_3$	134	CCCC(C)CC(C)CC	153	$C_1C(CC)CC_2(C)CC_{12}C$	165
$C_{12}(C)CCC(C_2)CC_1$	149	CCCCCCC(C)C)C	191.5	$C_1CCC_2(C)CC_{12}C(C)C$	169
$C_1CCCC_2C_1C_2C$	142	CC(C)CC(CC)CCC	190	$C_{12}CC_1CCC(C_2)C_3CC_3$	153
$C_1CCCC_2CC_{12}C$	202	CC(C)CC(C)CC(C)C	193	$C_1CC_2C_1C_3CCC_2CC_3$	155
$C_{12}CC_1CC(C_2)(C)C$	171	CCC(C)C(C)CCCC	182	$C_{12}CCC(CC_2)C_3C_1CCC_3$	170
$C_1CC_2(C)CC_1(C)C_2$	125.7	CCC(C)CC(CC)CC	190	$C_1(C)CC_2CCC_1C_3CC_{23}$	155.3
$C_{12}CCCC_1C_2(C)C$	117.6	CC(C)CCC(C)C(C)C	187	$C_1C(C_1(C)C_2CC_2)C_3CC_3$	162
$C_{12}C(C_2(C)C)C_1(C)C$	118.9	CC(C)C(CC)CCCC	178	$C_1CCCC_{12}C_3CCCC_{23}$	174
$C_{12}(C)CC_1(C)C_2$ (C)C	117.7	CCCC(C)C(C)CCC	173.5	$C_1CCCC_2(CC_2)C_{13}CC_3$	164
$C_{12}CC_1CC_3CC_3C_2$	109.1	CCCC(CCC)C(C)C	189.9	$C_{12}CCC(C_2C)C_3C_1C_3$	169.4
$C_{12}CC_1CCC_3C_2C_3$	118.5	CC(C)(C)CCCC(C)C	185	$C_1(C)(C)C_2C_3CC_1(C)CC_{23}$	159.3
$C_{12}CCC(C_2)C_3C_1C_3$	109.4	CCC(CC)C(C)CCC	182	$C_1C_2CC_3C_1(C)C_3C_2(C)C$	170.5
$C_1CC_2C_3CCC_2C_{13}$	115.6	CC(C)C(C)CC(C)CC	183	$C_{12}(C)C_3C_2CC(C_3)C_1(C)C$	168
$C_1(C)C_2CC_3C_1C_3C_2$	117.7	CCC(C)C(CC)CCC	183	$C_{12}CC_2C_3CC_1C_4(CC_4)C_3$	166
$C_{123}CC_1C_2C_4(C_3)CC_4$	106.8	CCC(C)(C)CCCCC	182	$C_1CC_2C_3CC_4C(C_3)C_{24}C_1$	188
$C_1C_2CC_3C_1(C)C_3C_2$	115.6	CC(C)CC(C)C(C)CC	178	$C_{12}CC_1C_3C_4C_2C_5C_3C_{45}$	167
$C_1CC_2(CC_2)C_{13}CC_3$	113.5	CC(C)(C)CCC(C)CC	174		
$C_{12}CC_3C_4CC_1C_2C_{34}$	112	CC(C)CC(CC)C(C)C	174.5		

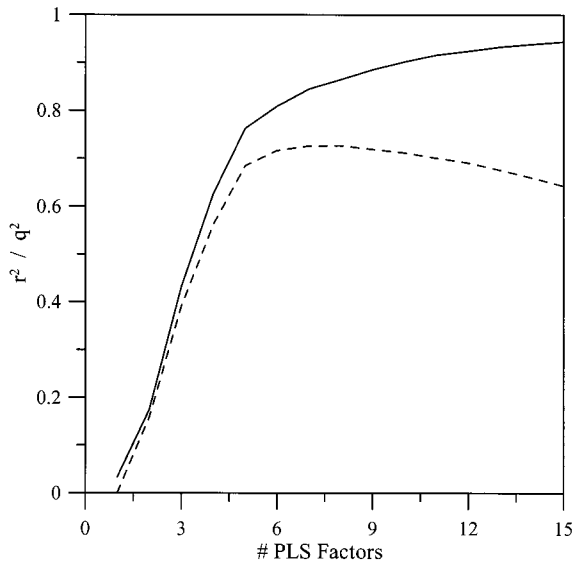

Figure 1. Evolution of r^2 (continuous line) and q^2 (dashed line) versus the number of PLS factors involved in the construction of the QSPR model for a set of 529 saturated hydrocarbons.

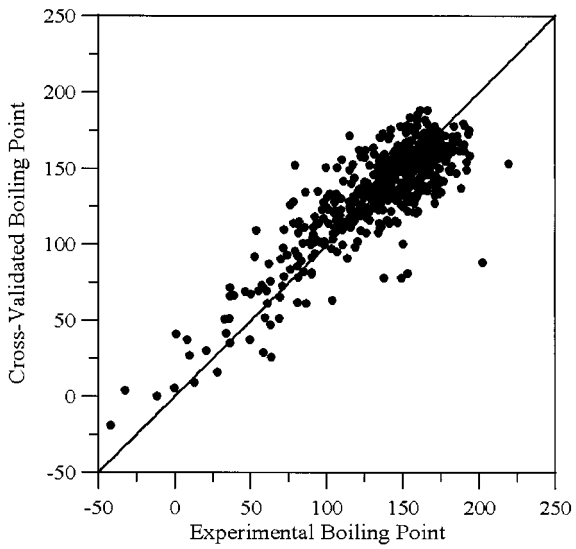

Figure 2. Experimental versus predicted boiling points for a set of 529 saturated hydrocarbons.

as the results provide the optimal balance between predictivity and descriptors applied, and a further increase does not lead to noticeable improved results and could lead to overfitted models. Thus, the final equation and statistical results for this molecular set are

$$\text{bp} = 3147.79\mathbf{f}_1 + 15269.10\mathbf{f}_2 + 34944.14\mathbf{f}_3 + 43295.57\mathbf{f}_4 + 47121.10\mathbf{f}_5 + 28552.70\mathbf{f}_6$$

$$r^2 = 0.808 \quad q^2 = 0.716 \quad s = 20.575$$

The results are also graphically presented by plotting the experimental bp versus the predicted ones from the PLS procedure, as shown in Figure 2. In addition, the random test was carried out permuting 100 times the bp values and reconstructing the QSPR model each time. The results are graphically compiled in Figure 3, where each point represents the values of r^2 and q^2 achieved in each permutation.

As seen from the statistical and graphical results, satisfactory results are obtained for this molecular set. Most of the compounds are correctly predicted, within a close margin, even few deviations are present, as displayed in Figure 2. The random test shows a clear separation between the original point (+) and the permuted ones (●). Among the permuted models, the highest values of r^2 and q^2 achieved are 0.475 and 0.029, respectively, whereas the average values

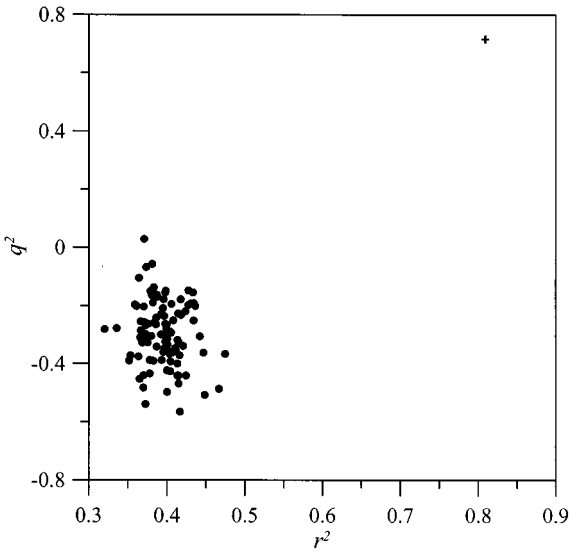

Figure 3. Random test results after permuting the boiling point values for a set of 529 saturated hydrocarbons 100 times. The real value is labeled with a cross.

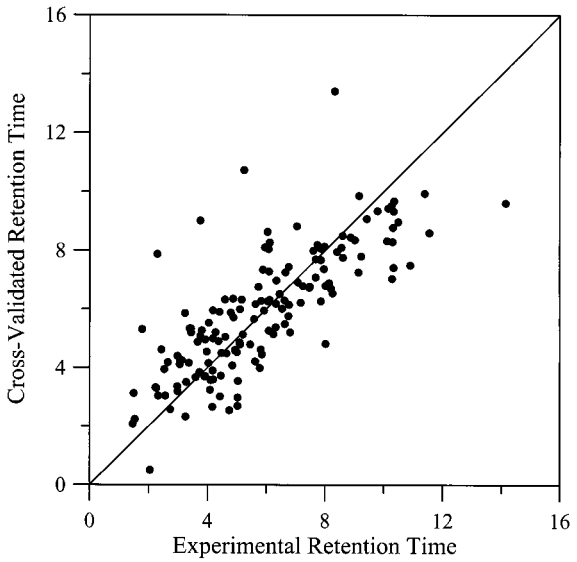

Figure 4. Experimental versus predicted retention times for a set of 152 compounds.

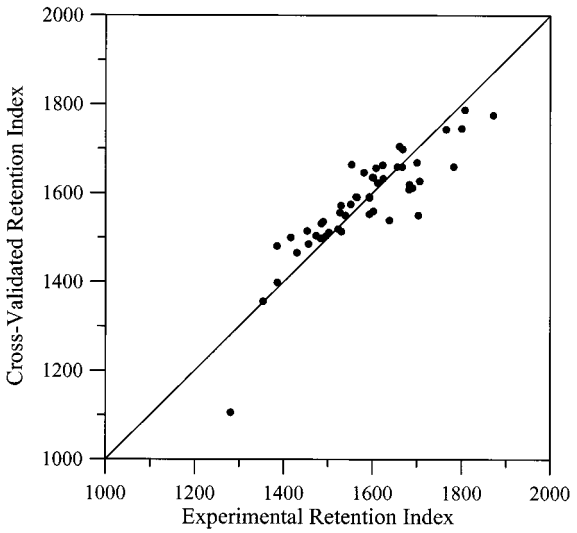

Figure 5. Experimental versus predicted retention indices for a set of 50 phenol derivatives.

are 0.395 and −0.293. The results of the previously published work[40] are slightly better, yielding to $r^2 = 0.961$ using 6 topological indices but providing no information about the predictive capacity of the proposed models.

Gas Chromatographic Retention Times for 152 Diverse Chemical Compounds. The identification of single organic compounds in chemical mixtures can be made by means of comparing chromatographic peaks with samples of each compound. Due to the costs, or simply by unavailability, of pure samples, sometimes this comparison cannot be done. In this way, it is interesting to develop QSPR models to describe and to predict retention parameters from molecular structures. The present example deals with retention times (t_R), namely the mean time needed for a molecule to elute from a chromatographic column. A molecular set composed by 152 structurally different compounds has been studied together with their retention times, presented in Table 2, with the expressed procedure. The original 152 × 152 descriptor matrix was reduced to 152 × 135 and submitted to the PLS routine. The optimal model was found to be built up from five PLS factors, and the obtained results are

$$t_R = 75.83\mathbf{f}_1 + 345.17\mathbf{f}_2 + 869.53\mathbf{f}_3 + 345.55\mathbf{f}_4 + 492.72\mathbf{f}_5$$

$$r^2 = 0.753 \quad q^2 = 0.541 \quad s = 1.657$$

As seen from the statistical results, an acceptable relationship with valuable predictive capacity is obtained. The predictive power is also graphically represented in Figure 4, where experimental versus predicted t_R values are plotted.

As seen in Figure 4, all compounds are accurately predicted in a narrow margin with the exception of a few molecules, which are mostly overestimated. The results of the random test for this molecular set are, similarly to the previous examples, very satisfactory, yielding to average r^2 and q^2 values of 0.229 and −0.234, whereas the maximum values achieved were 0.373 and 0.180, respectively. The previous work[41] obtained better results using a large set of 296 classical, topological, and quantum parameters computed with the CODESSA[45] software. Their models were calculated using linear and nonlinear multilinear regressions yielding to cross-validation regression coefficients (r^2_{cv}) of 0.961 and 0.967, respectively. Excellent results in the random tests were also reported.

Gas−Liquid Chromatography Retention Indices for 50 Phenol Derivatives. This example, similarly to the previous one, deals with chromatography. However, instead of t_R, Kovátz retention indices (RI)[46] are studied. The analyzed molecular set consists of 50 phenol derivatives. The molecular structures, as well as the RI, are presented in Table 3. The original 50 × 50 descriptor matrix was reduced to 50 × 38 after the redundancies elimination, and the optimal QSPR model, built up from 6 PLS factors, results in

$$\mathrm{RI} = 12938.48\mathbf{f}_1 + 31183.31\mathbf{f}_2 + 54224.34\mathbf{f}_3 + 60530.94\mathbf{f}_4 + 25264.86\mathbf{f}_5 + 45970.82\mathbf{f}_6$$

$$r^2 = 0.897 \quad q^2 = 0.762 \quad s = 59.294$$

Table 2. SMILES Notation[47,48] and Retention Times of 152 Diverse Chemical Compounds[41]

compd	retention time	compd	retention time	compd	retention time
CCCCCCCCCCCC	6.34	$c_1cm_3cc(c_1)Oc_2cm_3ccc_2$	7.74	$c_1cm_3cc(O)c_1OC(C)C$	5.95
CCCCCCCCCCCC CCC	8.56	$c_1cm_3cc(c_1)C\#N$	4.1	$c_1cnccc_1C\#N$	3.89
CCCCCCC=C	2.53	$c_1cm_3cc(c_1)Sc_2cm_3ccc_2$	9.02	$c_1(CC)cccnc_1$	3.92
CCCCCCCCC=C	4.45	$c_{12}CCCCc_1c(O)ccc_2$	8.02	$c_1c(C)cc(C)nc_1C$	4.37
$C_1CCC(O)CC_1$	3.27	$c_1cm_3ccc_1C(=O)C_2CCCCC_2$	9.14	$c_1c(C)cc(C)nc_1N$	5.85
$C_1CCC(=O)CC_1$	3.25	$c_1cm_3cc(c_1)NC$	4.94	$c_1cm_3cc(c_1)C=CC=O$	6.63
CCCCCCCCN	4.88	$c_{12}cm_3cc(O)c_2cm_3c(c_1)O$	10.33	$c_{12}cm_3ccc_1OC(=O)CC_2$	7.46
CCCCCCCCCCN	6.62	$C_1=CC=CS_1$	1.49	$c_1cm_3cc(c_1)C=CC(=O)O$	7.86
CCCCCC#N	3.06	$C_1=CC=CN_1$	2.04	$c_{12}cm_3ccc_1OC(=O)C=C_2$	7.87
$C_1CCC(C_1)=O$	2.26	$c_1(CN)ccccc_1$	4.47	$c_1(CC)ccncc_1$	3.97
CCCCCCCCSCCCCCCCCC	10.9	$c_1(C)ccc(cc_1)n(o)o$	6.24	$c_1(CCN)ccncc_1$	6.08
CCOC(=O)OCC	2.29	$c_1cm_3c(cc_1)C(=O)Oc_2$ cm_3ccc_2	9.42	$c_1CCN(CC_1)C(=O)OCC$	6.11
$C_1CCC(C_1)O$	2.32	$c_1(C)ccc(N)cc_1$	5.02	$c_1cm_3nc(c_1)CC$	3.44
CCCCCCCCO	5.1	$c_{12}cm_3ccc_1Cc_3c_2cm_3cc_3$	10.28	$c_1cm_3nc(c_1)CCO$	5.63
CCCCCCCCCCO	6.77	$c_1(O)ccc(cc_1)CC$	5.81	$c_1cm_3c(cc_1)C(=O)C(=O)c_2cm_3ccc_2$	10.31
CCCCCCCCCCCCO	8.26	$c_1(N)ccccc_1$	4.08	$c_1(ccncc_1)C(=O)OCC$	6.04
$C_1CCCC(N)CC_1$	4.4	$c_1(N)ccccc_1C$	5	$c_1cm_3nc(c_1)C(C)=O$	4.6
$C_1CCC(N)CC_1$	3.05	$c_1cm_3ccc_1OCCCC$	6.08	$c_1cncc(c_1)C(C)=O$	5.2
CCCCNCCCC	4.18	$c_1cm_3ccc_1Br$	3.75	$c_1cm_3c_2CCOCc_2c_1$	6.04
CCCCCCCCCCCN	8.14	$c_1cm_3ccc_1I$	4.74	$C_1(CN)=CC=CO_1$	2.73
CCCCCCCCCCBr	7.42	$c_1cm_3ccc_1C=O$	3.92	$C_1=CC=CN(C)C_1=O$	5.78
$N_1(C)CCNCC_1$	2.98	$c_1cm_3ccc_1CO$	4.64	$c_1cm_3c_2c(c_1)cnc_3cm_3ccc_{23}$	10.49
C(O)CCO	2.65	$c_1cm_3(C)ccc_1C=O$	5.1	$c_1cm_3c(CC)c(c_1)O$	5.58
C(O)CCCO	3.72	$C_1(=O)C=CC(=O)c_2c_1cm_3cc_2$	7.68	$c_1cm_3c(cn_1)C=O$	4.17
$C_1CCC(C_l)CC_1$	3.36	$c_{12}cm_3ccc_1C=CN_2$	6.81	$c_1cm_3c_2cm_{33}$ cm3ccc3nc2c_1	10.35
$C_1CCCC(=O)CC_1$	4.42	$c_1(ccccc_1)n(o)o$	5.09	$c_1cm_3c(cc_1)Nc_2cm_3ccc_2$	9.16
$C_1CCC(CC_1)C(=O)O$	5.62	$c_{12}cm_3ccc_1N=CS_2$	6.33	$c_1(C)ccc(O)cc_1$	5.02
$c_1cm_3cc_2cm_3ccc_{12}$	6.09	$c_{12}cm_3ccc_1NC=N_2$	8.02	$c_1(C)ccccn_1$	2.56
$c_1cm_3cc(c_1)CCc_2cm_3ccc_2$	7.6	$c_{12}cm_3ccc_1cm_3c(n_2)C$	7.04	$c_1cm_3cc(c_1)Cc_2cm_3ccc_2$	7.96
$c_1cm_3cc(c_1)CCC$	4.04	$c_1(C)ncccc_1C$	3.8	$c_1cm_3cc(c_1)CCc_2cm_3ccc_2$	8.61
$c_1cm_3cc(c_1)CC=C$	3.39	$c_1(C)cccc(n_1)C$	3.23	$c_1cm_3c(cc_1)CC(C)(C)Cc_2cm_3ccc_2$	9.79
$c_1cm_3cc_2CCCc_{12}$	4.79	$c_1(ccccc_1)SSc_2cm_3ccc_2$	10.15	$c_1cm_3(C)ccc_1Cl$	4.03
$c_1cm_3cc(c_1)C(C)C$	3.75	$c_1(C)nccc(c_1)C$	3.66	$c_1cm_3cc(c_1)CBr$	5.24
$c_1cm_3cc_2c_1cm_3c_3cm_3ccc_{23}$	10.27	$c_1(C)ccccc_1O$	4.84	CC(C)C(=O)C(C)C	2.43
$c_1cm_3c_2CCCCc_2c_1$	5.92	$c_1(OC)cc(OC)cc(c_1)OC$	7.68	$c_1cm_3ccc_1$	1.46
$c_1cm_3c_2CCCNc_2c_1$	7.17	$c_1(O)ccc(cc_1)C(C)C$	6.32	$c_1cm_3ccc_1C$	2.23
$c_1cm_3c_2cm_3cnc_2c_1$	6.45	$c_1cm_3cc(c_1)C(=O)c_2cm_3ccc_2$	9.23	$c_1cm_3cc(C)c_1C$	3.43
$c_1c(C)cc(C)cc_1C$	4.18	$CC(=O)c_1cm_3ccc_1$	4.93	C(CC)CCCCCCC	4.59
$c_1cm_3c_2CNCCc_2c_1$	6.75	CCCCCCCCC(=O)O	6.76	C(CC)CCCCCCCCCCCCCCCCC	11.55
$c_1cm_3c_2cm_{33}cm_3ccc_3cm_{32}c_1$	10.34	$c_1cm_3ccc_1OC(=O)C$	5.73	C_1CCCCC_1	1.53
$c_1cm_3c(cc_1)C(=C)C$	4.27	$c_1(O)ccccc_1C(C)C$	6.11	CCCCCCC	1.78
$c_1cm_3c_2c(c_1)Oc_3cm_3ccc_{23}$	8.6	$C(C\#N)c_1cm_3ccc_1$	5.45	$c_1cm_3ccc_1CC$	3.12
$c_1cm_3c_2c(c_1)Sc_3cm_3ccc_{23}$	10.11	$C(C(=O)O)c_1cm_3ccc_1$	6.54	$c_1cm_3ccc_1C(=O)CC$	5.82
$c_1(cccc_2cm_3ccc_{12})c_3cm_3cc_4cm_3ccc_{34}$	14.14	$c_1cm_3cc(c_1)CCCC(=O)O$	7.86	$c_1cm_3ccc_1OC$	3.59
$c_1cm_3cc_2CCCC(=O)c_{12}$	7.48	$c_1cm_3cc(c_1)SC$	5.16	$c_1(O)ccc_2cm_3cc(O)c_2c_1$	10.3
$c_1cm_3cc_2cm_3cc(Br)c_{12}$	8.33	$c_1cm_3cc(c_1)CSC$	5.88	$c_1cm_3c_2cm_3ncc_2c_1$	6.65
$c_1cm_3cc_2cm_3cc(O)c_{12}$	8.41	$c_1nccc(c_1)C$	2.98	$c_1cm_3cnc_1c_2ncccc_2$	7.98
$c_1cm_3ccc_1C_2CCCCC_2$	7.24	c_1ncccc_1C	2.97	$C_1(C)cc(C)cc(C)c_1C(O)=O$	7.82
$c_1cm_3c_2cm_3cc(Cc_3cm_3ccc_3)c_2c_1$	11.39	$c_1c(C)cccc_1O$	5.03	$c_1cm_3ccc_1S(C)=O$	7.07
$c_1cm_3cc(c_1)O$	4.16	$c_1(C)c(C)c(C)c(C)c(C)c_1C$	8.21	$c_1cm_3ccc_1CCC=C$	4.87
$c_1cm_3cc(c_1)OCc_2cm_3ccc_2$	8.87	$n_1cm_3cc(c_1)C\#N$	4.19		

As seen from the previous statistical results, a satisfactory QSPR model is found with a remarkable predictive capacity, as judged from the q^2 value achieved. A clearer picture of the obtained results can be seen by browsing Figure 5, where the experimental RI are plotted versus the cross-validated ones. As seen in Figure 5, all RI are correctly predicted except the one referred to the unsubstituted benzene, which is underestimated. In addition, the random test is satisfactorily surpassed, yielding to averaged values of 0.437 and −0.352 for r^2 and q^2, respectively, whereas the maximum values were 0.675 and 0.377.

The previous work[46] achieved valuable results with a few number of molecular descriptors, consisting of vertex- and edge-weighted molecular graphs. Results ranging from 0.978 to 0.982 for r^2_{cv} values are obtained with up to 4 topological descriptors.

CONCLUSIONS

A number of molecular sets have been tested for correlation search with molecular properties, such as boiling points and chromatographic parameters. Satisfactory correlations are obtained using a small number of descriptors for molecular sets making use of quantum similarity measures.

Even if other methodologies may provide better results, it must be emphasized that the procedure used along this work, as well as the generation of molecular descriptors, has been kept unmanipulated. In other words, the exposed QSPR protocol consists of a general methodology made of unbiased and universal MQSM descriptors able to characterize different molecular properties without introducing further information. Additional refinements may be applied to the procedure to improve the results according to each

Table 3. Smiles Notation[47,48] and Retention Indices of 50 Phenol Derivatives[42]

compd	retention index	compd	retention index
$c_1(O)ccccc_1$	1281	$c_1(O)c(C)ccc(C)c_1C$	1551
$c_1(O)ccccc_1C$	1354	$c_1(O)cc(C)c(C)cc_1C$	1593
$c_1(O)cccc(c_1)C$	1386	$c_1(O)cc(C)c(C)c(c_1)C$	1667
$c_1(O)ccc(C)cc_1$	1385	$c_1(O)ccc(cc_1)CC(C)C$	1612
$c_1(O)ccccc_1CC$	1430	$c_1(O)ccccc_1CCCC$	1600
$c_1(O)cccc(c_1)CC$	1483	$c_1(O)cccc(c_1)CCCC$	1668
$c_1(O)ccc(cc_1)CC$	1473	$c_1(O)ccc(cc_1)CCCC$	1661
$c_1(O)cccc(C)c_1C$	1495	$c_1(O)ccc(cc_1C)CCC$	1623
$c_1(O)ccc(C)cc_1C$	1456	$c_1(O)c(C)cccc_1CCC$	1553
$c_1(O)cc(C)ccc_1C$	1453	$c_1(O)cc(C)ccc_1CCC$	1602
$c_1(O)c(C)cccc_1C$	1416	$c_1(O)ccc(C)cc_1CCC$	1593
$c_1(O)cc(C)cc(c_1)C$	1489	$c_1(O)ccc(CC)cc_1CC$	1602
$c_1(O)ccc(C)c(c_1)C$	1530	$c_1(O)cc(CC)ccc_1CC$	1624
$c_1(O)ccc(cc_1)C(C)C$	1527	$c_1(O)ccc(CC)c(c_1)CC$	1682
$c_1(O)ccccc_1CCC$	1502	$c_1(O)cc(C)c(C)c(C)c_1C$	1782
$c_1(O)cccc(c_1)CCC$	1565	$c_1(O)c(C)cc(C)c(C)c_1C$	1690
$c_1(O)ccc(cc_1)CCC$	1563	$c_1(O)c(C)c(C)cc(C)c_1C$	1683
$c_1(O)ccc(C)cc_1CC$	1523	$c_1(O)cc(C)c(C)cc_1CC$	1656
$c_1(O)cc(C)ccc_1CC$	1529	$c_1(O)ccccc_1CCCC$	1700
$c_1(O)c(C)cccc_1CC$	1485	$c_1(O)ccc(cc_1)CCCCC$	1765
$c_1(O)cc(C)cc(c_1)CC$	1581	$c_1(O)ccc(cc_1)CC(C)(C)C$	1703
$c_1(O)ccc(CC)cc_1C$	1539	$c_1(O)c(CC)cccc_1CCC$	1706
$c_1(O)ccc(CC)c(c_1)C$	1608	$c_1(O)ccccc_1CCCCC$	1800
$c_1(O)ccc(C)c(C)c_1C$	1638	$c_1(O)ccc(cc_1)CCCCCC$	1871
$c_1(O)cc(C)cc(C)c_1C$	1593	$c_1(O)cc(CCCC)ccc_1CC$	1807

molecular set under study; however, the exposed protocol may constitute an excellent starting point for subsequent research.

ACKNOWLEDGMENT

This research has been partially supported by Project No. SAF2000-0223-C03-01 from the Spanish Ministerio de Ciencia y Tecnología. X.G. benefits from a predoctoral fellowship from the University of Girona. Financial support from the Fundació María Francisca de Roviralta is also acknowledged.

REFERENCES AND NOTES

(1) Quoted in: Borman, S. New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News* **1990**, *68*, 20−23.
(2) Hammett, L. P. The Effect of Structures upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96−103.
(3) Jurs, P. C. Quantitative Structure−Property Relationships (QSPR). In *Encyclopedia of Computational Chemistry*; Schleyer, P.v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III; Schreiner, P. R., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1998; Vol. 4, pp 2320−2330.
(4) Waterbeemd, H. v. d., Ed. *Structure−Property Correlations in Drug Research*; Academic R. G. Landes Co.: Austin, TX, 1996.
(5) Kubinyi, H. Quantitative Structure−Activity Relationships in Drug Design. In *Encyclopedia of Computational Chemistry*; Schleyer, P.v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III; Schreiner, P. R., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1998; Vol. 4, pp 2309−2319.
(6) Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(7) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *THEOCHEM* **1998**, *451*, 11−23.
(8) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 1−42.
(9) Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets; vector semispaces and convex sets; quantum similarity measures and ASA density functions; diagonal vector spaces and quantum chemistry. In *Advances in molecular similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 43−72.
(10) Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and recent developments on molecular quantum similarity. *Top. Curr. Chem.* **1995**, *173*, 31−62.
(11) Carbó, R.; Besalú, E. Theoretical foundations of quantum molecular similarity. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches*; Carbo, R., Ed.; Kluwer: Amsterdam, 1995; pp 3−30.
(12) Carbó-Dorca R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures: concepts, definitions and applications to quantitative structure-properties relationships. In *Advances in molecular similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press: Greenwich, CT, 1996; Vol. 1, pp 1−42.
(13) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships. *J. Math. Chem.* **1995**, *18*, 237−246.
(14) Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25−32.
(15) Lobato, M.; Amat, L.; Carbó-Dorca, R. Structure−activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, *16*, 465−472.
(16) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: an antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624−631.
(17) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure−activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(18) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 597−610.
(19) Gironés, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron−electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 477−485.
(20) Robert, D.; Carbó-Dorca, R., Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401−422.
(21) Gironés, X.; Amat, L.; Carbó-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999**, *10*, 545−556.
(22) Dean, P. M. Molecular Similarity. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers BV.: Leiden, The Netherlands, 1993; pp 150−172.
(23) Richards, W. G., Molecular Similarity and Dissimilarity. In *Modeling of Biomolecular Structures and Mechanisms*; Pullman, A., Jortner, J., Pullman, B., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 365−369.
(24) Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA: a molecular superposition program based on Topo-Geometrical Considerations. *J. Comput. Chem.* **2001**, *22*, 255−263.
(25) Carbó-Dorca, R. Stochastic Transformation of Quantum Similarity Matrixes and Their Use in Quantum QSAR (QQSAR) Models. *Int. J. Quantum. Chem.* **2000**, *79*, 163−177.
(26) Constants, P.; Carbó, R. Atomic shell approximation: electron density fitting algorithm restricting coefficients to positive values. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1046−1053.
(27) Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation: first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem.* **1997**, *18*, 2023−2039.
(28) Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diamminedichloroplatinum(II) complex as an application example. *J. Comput. Chem.* **1999**, *20*, 911−920.
(29) *AMPAC 6.55*; Semichem: Shawnee, KS, 1999.
(30) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Chemical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(31) Höskuldsson A. *Prediction Methods in Science and Technology*; Thor: Copenhagen, 1996.
(32) Tenenhaus, M. *Regression de PLS*; Editions Technip: Paris, 1997.
(33) Wold, S.; Johansson, E.; Cocchi, M. PLS−Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Desing, Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science: Leiden, The Netherlands, 1993; pp 523−550.
(34) Wold, S. PLS for multivariate linear modelling. In *Methods and Principles in Medicinal Chemistry. Vol. 2*: *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195−218.

(35) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect on Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(36) Wold, S.; Sjöström, M.; Eriksson, L. Partial Least Squares Projections to Latent Structures (PLS) in Chemistry. In *Encyclopedia of Computational Chemistry*; Schleyer, P v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III., Screiner, P. R., Eds.; John Wiley and Sons Ltd.: Chichester, U.K., 1994; Vol. 4, pp 2006−2021.
(37) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Turorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(38) Montgomery, D. C.; Peck, E. A. *Introduction to linear regression analysis*; Wiley: New York, 1992.
(39) Wold, S. Cross-validatory estimation of a number of components in factor and principal component models. *Technometrics* **1978**, *20*, 397−405.
(40) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788−802.
(41) Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.
(42) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. Comparison of Weighting Schemes for Molecular Graph Descriptors: Application in Quantitative Structure-Retention Relationship Models for Alkylphenols in Gas−Liquid Chromatography. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732−743.
(43) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28−41.
(44) Ivanciuc, T.; Balaban, A. T. Quantitative Structure−Property Relationship Study of Normal Boiling Points for Halogen-/Oxygen-/Sulphur-Containing Organic Compounds Using the CODESSA Program. *Tetrahedron* **1998**, *54*, 9129−9142.
(45) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.
(46) Buryan, P.; Nabivach, V. M., Dimitrikov, V. P. Structure-Retention Correlations of Isomeric Alkylphenols in Gas−Liquid Chromatography. *J. Chromatogr.* **1990**, *509*, 3−14.
(47) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(48) SMILES stands for simplified molecular input line entry system. More information can be found at the following: http://esc.syrres.com/interkow/docsmile.html.

CI0103370