Prediction of Cellular Toxicity of Halocarbons from Computed Chemodescriptors: A Hierarchical QSAR Approach

Subhash C. Basak,*,[†] Krishnan Balasubramanian,[‡,§,||] Brian D. Gute,[†] Denise Mills,[†]
Anna Gorczynska,[#] and Szczepan Roszak[#]

Natural Resources Research Institute, University of Minnesota−Duluth, 5013 Miller Trunk Highway,
Duluth, Minnesota 55811, Department of Applied Sciences, University of California Davis, Hertz Hall,
L-794 Livermore, California 94550, Chemistry & Material Science Directorate, Lawrence Livermore National
Laboratory, Livermore, California 94550, Glenn T. Seaborg Center, Lawrence Berkeley National Laboratory,
Berkeley, California 94720, and Institute of Physical and Theoretical Chemistry, Wroclaw University of
Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland

Received September 10, 2002

A hierarchical quantitative structure−activity relationship (HiQSAR) approach was used to estimate toxicity and genetic toxicity for a set of 55 halocarbons using computed chemodescriptors. The descriptors consisted of topostructural (TS), topochemical (TC), geometrical, semiempirical (AM1) quantum chemical, and ab initio (STO-3G, 6-31G(d), 6-311G, 6-311G(d), and aug-cc-pVTZ) quantum chemical indices. For the two toxicity endpoints investigated, ARR and D_{37}, the TC indices gave the best cross-validated R^2 values. The 3-D indices also performed either as well as or slightly superior to the TC indices. For the four categories of quantum chemical indices used for the development of predictive models, the AM1 parameters gave the worst performance, and the most advanced ab initio (B3LYP/aug-CC-pVTZ) parameters gave the best results when used alone. This was also the case when the quantum chemical indices were used in the hierarchical QSAR approach for both of the toxicity endpoints, ARR and D_{37}. The models resulting from HiQSAR are of sufficiently good quality to estimate toxicity of halocarbons from structure.

1. INTRODUCTION

An important objective of contemporary computational toxicology is the prediction of the potential toxic effects of chemicals from their computed molecular descriptors.[1−20] Thousands of chemicals are already in the marketplace, and their numbers are growing at an ever increasing rate. The TSCA Inventory of USEPA currently has over 81 000 entries, and the list is increasing yearly.[21] Most of these chemicals have none of the experimental data necessary for the estimation of their effects on human and environmental health.[22] In the areas of chemical ecotoxicity and genotoxicity, the availability of data is even more appalling. Only 15% of TSCA chemicals have genotoxicity data; less than 6% have chronic toxicity or ecotoxicity data.[22] Only a few hundred of the many thousands of candidate chemicals have been tested in the National Toxicology Program (NTP) for carcinogenicity in the 2-year rodent bioassay program.

It is clear from the above that, in the foreseeable future, hazard assessment of chemicals must be carried out in a data-poor situation, while the number of candidate chemicals is increasing at a rate much faster than the availability of experimental physicochemical and toxicity data prerequisite to the proper hazard estimation of all these substances. A viable solution to this quagmire is the use of quantitative structure−activity relationship (QSAR) models which use algorithmically derived properties, i.e., properties which can be calculated directly from chemical structure without the input of any further experimental data.

Halocarbons constitute an important class of chemicals as solvents and useful intermediates in various synthetic processes. So, the potential of their release to the environment worldwide is substantial.[23−25] Therefore, it was of interest to develop QSARs of halocarbons using descriptors that can be calculated from structure. Crebelli et al. determined various toxicity endpoints for a set of 55 halocarbons.[26] They formulated QSAR models using a mixture of calculated and experimental properties, viz., molecular refractivity (MR), octanol/water partition coefficient (log*P*), and quantum chemical indices such as HOMO energy, LUMO energy, and the difference between HOMO and LUMO energies calculated by ab initio methods. While MR characterized the generalized shape and size of molecules, the quantum chemical indices quantify the electrophilic nature of the molecules. In our hierarchical QSAR (HiQSAR) approach, we have used a combination of chemodescriptors, calculated directly from molecular structure, for the formulation of QSARs to predict various toxicity endpoints.[5−8,11−13,27−31] The HiQSAR method begins model building with parameters which can be computed most easily; parameters demanding more computational resources are added only if the easily calculable indices do not give satisfactory results. We have routinely used topostructural (TS), topochemical (TC), geometrical, and semiempirical quantum chemical indices for the development of QSARs for toxicity estimation. Previous

[*] Corresponding author phone: (218)720-4230; fax: (218)720-4328; e-mail: sbasak@nrri.umn.edu.
[†] University of Minnesota−Duluth.
[‡] University of California Davis.
[§] Lawrence Livermore National Laboratory.
[||] Lawrence Berkeley National Laboratory.
[#] Wroclaw University of Technology.

10.1021/ci020054n CCC: $25.00
Published on Web 05/06/2003

results with halocarbons by Crebelli et al. showed that they achieved moderately good results in QSAR development using a combination of calculated descriptors.[26] So, in the current study we have attempted to develop QSARs for the set of 55 halocarbons using a combination of topological, geometrical, and quantum chemical indices using the HiQSAR approach.

2. METHODS AND MATERIALS

2.1. Halogenated Aliphatic Hydrocarbon Database. The database of halogenated aliphatic hydrocarbons used in this study was taken from the work of Crebelli et al.[26] Crebelli and co-workers present test data for mitotic chromosome malsegregation and lipid peroxidation tested in *Aspergillus nidulans* diploid strain P1. Results presented for mitotic chromosome malsegregation include lowest effective concentration affecting chromosome distribution (LEC), lowest concentration for arresting conidial development (ARR), and lowest concentration inducing one lethal hit per cell (D_{37}). Lipid peroxidation study results were presented as lowest effective concentration resulting in lipid peroxidation (LEC_{LP}); however, not all of the chemicals were tested for lipid peroxidation. For the purposes of this study, we chose to model the two continuous biological endpoints (ARR and D_{37}) that were available for all 55 chemicals. It should be noted that LEC data were available for all of the chemicals in the set as well; however, activity values were only available for 24 of the 55 chemicals, the remaining 31 chemicals tested negative. Chemical names and activity data (ARR and D_{37}) are provided as Table S1. Figure S1 shows the optimized geometries of all 55 neutral halocarbons in the Crebelli set considered in this study. These geometries were optimized using the using the high-level cc-pVTZ basis set using the DFT/B3LYP method. The basis set and the technique are considered quite reliable for the geometry optimization. After the geometry optimization of the neutral halocarbons we computed the energies of the anions at the optimized neutral geometries. The vertical electron affinities are taken as the difference of the geometries of the anion and neutral species at the optimized neutral geometries. For further details of the testing regimens, please refer to the original work of Crebelli and co-workers.[26,32,33]

2.2. Calculation of Topological Indices. The topological indices (TIs) used in this study were calculated using three main software programs: POLLY 2.3,[34] MolConn-Z 3.50,[35] and Triplet.[36] Included in the suite of 192 topological indices used in this study are the following: Wiener number,[37] molecular connectivity indices as calculated by Randić[38] and Kier and Hall,[39] frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić[40] as well as those of Raychaudhury et al.,[41] parameters defined on the neighborhood complexity of vertices in hydrogen-filled molecular graphs,[42−44] Balaban's *J* indices,[45−47] local orthogonal vertex invariants,[36] and the electrotopological indices of Kier and Hall.[48] More information on the topological indices calculated by POLLY has been reported in earlier studies.[4,10,13,49]

2.3. Calculation of Geometrical Indices. The geometrical indices include van der Waals volume,[50] the three-dimensional Wiener numbers for both the hydrogen-filled and hydrogen-suppressed molecular structures,[51] and the Kappa shape descriptors of Kier and Hall.[52,53] van der Waals volume, V_W, was calculated using *SYBYL 6.4* from Tripos Associates, Inc.[54] The 3-D Wiener numbers were calculated in the *SYBYL* interface using an SPL (Sybyl Programming Language) program developed in our lab. Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using *CONCORD 3.2.1.*[55] Finally, the Kappa shape descriptors were calculated using MolConn-Z 3.50.[35]

2.4. Quantum Chemical Parameters. Semiempirical quantum chemical parameters were calculated using the Austin Model version one (AM1) Hamiltonian. These parameters were calculated using *MOPAC 6.00* in the *SYBYL* interface.[56] Ab initio quantum chemical calculations were performed applying the second-order Møller−Plesset perturbation theory[57] for the sequence of standard Pople's atomic basis sets: STO-3G,[58,59] 6-31G(d),[60−65] 6-311G,[66,67] 6-311G-(d), and aug-cc-pVTZ.[68−72] Additional calculations were done within the density functional theory (DFT) approach.[73] The DFT calculations utilized Becke's three-parameter functional[74] with Vosco et al.'s[75] local correlation part and Lee et al.'s[76] nonlocal part (abbreviated as B3LYP). DFT studies were done applying extended aug-cc-pVTZ basis set.[57,73−76] This basis set consists of a set of diffuse functions to account for the negative electronic charge distribution on anions. The density functional theory has been shown as a reliable method for reproducing electron affinity properties.[77]

The calculations in the MP2/STO-3G approach were utilized as the lowest-level ab initio scheme, similar to semiempirical quantum chemical approaches. All calculations applying Pople's basis sets were done at the MP2 level of theory and are denoted in the text by the name of the atomic basis sets. The DFT calculations are referred to as B3LYP/aug-cc-pVTZ. The results from the ab initio calculations are presented in Tables S2−S5. Table S5 provides additionally experimental dipole moments[17] for the assessment of the quality of the charge distribution. All ab initio quantum chemical calculations were conducted using Gaussian98W.[78]

2.5. Statistical Analysis. Initially, the majority of topological indices were transformed by the natural logarithm of the index plus one. This was done to scale the indices, since some may be several orders of magnitude greater than others, while other indices may equal zero. The geometric indices were transformed by the natural logarithm of the index for consistency, the addition of one was unnecessary. Other indices, including the majority of the quantum chemical indices, were also transformed using natural logarithm; however, constants were chosen on a case-by-case basis to bring the minimal value for the descriptor above zero. Finally, the ab initio values for $Gap_{HOMO-LUMO}$ and VEA were not scaled since their values were already distributed between negative one and one.

The set of 192 topological indices was then partitioned into the two distinct sets: topostructural (TS) indices (86) and topochemical (TC) indices (106). TS indices are topological indices that encode information about the adjacency and distances of atoms (vertices) in molecular structures (graphs) irrespective of the chemical nature of the atoms involved in the bonding or factors such as hybridization states of atoms and number of core/valence electrons in individual atoms. TC indices are parameters that quantify information regarding the topology (connectivity of atoms) as well as specific chemical properties of the atoms and bonds comprising a molecule. TC indices are derived from weighted molecular graphs where each vertex (atom) is properly weighted with relevant chemical/physical properties. Each of the sets, TS and TC indices, were divided into subsets, or principal components (PCs), using the SAS principal components analysis (PCA) procedure (PRINCOMP)[79] to further reduce the number of independent variables for use in model construction. This procedure creates a new set of completely orthogonal PCs derived from linear combinations of all indices. Only PCs with eigenvalues greater than or equal to one have been retained for this study. A more detailed explanation of this approach has been provided in a previous study by Basak et al.[49] After the PCA, a correlation analysis was conducted on the PCs to determine which TIs were most highly correlated with each of the PCs. The TI most-highly correlated with each PC was then selected for modeling, resulting in a greatly reduced set of indices. The PCA and selection of indices was performed independently for both the topostructural and topochemical indices and resulted in a set of five topostructural indices and a set of 12 topochemical indices.

As is well-known, the smaller the data set, the greater the chance of spurious error when using a large number of independent variables (descriptors), so reducing the number of independent variables is crucial when modeling of data set of this size. Topliss and Edwards[80] have thoroughly studied this issue of chance correlations. For a set with between 50 and 60 dependent variables (observations), to keep the probability of chance correlations less than 0.01, at most about 45 independent variables may be used (if a correlation coefficient of $R^2 \geq 0.9$ is achieved). This number is dependent on the actual correlation achieved in the modeling process, a higher correlation results in a better chance of using more variables with the same limited probability of chance correlations. In this study we are well below the cutoff of 45 independent variables. In fact, the total number of descriptors which will be considered for all models and estimation is 23 (this includes the topostructural, topochemical, and geometrical indices), well within the bounds of the Topliss and Edwards criteria.[80] Addition of all quantum chemical indices employed in hierarchical modeling takes the total up to 60, beyond the allowable bounds of the Topliss and Edwards criteria. However, we consider the descriptors on the quantum chemical level to be exclusive, in other words we never use parameters from the four different quantum chemical approaches at the same time. So for any HiQSAR including quantum chemical parameters, the total number of indices considered in modeling never exceeds 30, again well within the bounds of the Topliss and Edwards criteria.

Regression modeling was accomplished using the SAS procedure REG[79] on nine distinct sets of indices. These sets were constructed as part of the HiQSAR approach to model development. The hierarchy begins with the simplest parameters, the TS indices. After using the TS indices to model the activity, the next level of parameters of higher complexity are added. To the indices included in the best TS index model, we add all of the TC indices and proceed to model the activity using these parameters. Likewise, the indices included in the best model from this procedure are combined with the indices from the next complexity level, the geometrical indices and modeling is conducted once again. Finally, the best model utilizing TS indices, TC indices, and geometrical indices is combined with the various quantum chemical parameters to develop the highest level models in the hierarchy.

Table 1. Five Principal Components Retained from the PCA of 86 Topostructural Indices

PC	eigenvalue	proportion of variance	cumulative variance	most-highly correlated TSI
1	54.2983	0.6314	0.6314	W
2	21.2275	0.2468	0.8782	$DS1_3$
3	5.4819	0.0637	0.9420	$^4\chi_{PC}$
4	2.2830	0.0265	0.9685	$^6\chi_C$
5	1.1342	0.0132	0.9817	$^5\chi_{PC}$

Table 2. 12 Principal Components Retained from the PCA of 106 Topochemical Indices

PC	eigenvalue	proportion of variance	cumulative variance	most-highly correlated TCI
1	37.9045	0.3384	0.3384	AZV_3
2	20.7565	0.1853	0.5238	$^2\chi^b$
3	13.1183	0.1171	0.6409	Q_V
4	9.7944	0.0875	0.7283	$^1\chi^v$
5	6.7769	0.0605	0.7888	$^4\chi^v_{PC}$
6	4.9909	0.0446	0.8334	SdssC
7	3.6443	0.0325	0.8659	SdsCH
8	2.4341	0.0217	0.8877	$SdCH_2$
9	2.1676	0.0194	0.9070	$^5\chi^v_{PC}$
10	1.6602	0.0148	0.9219	SssssC
11	1.4927	0.0133	0.9352	H_{MIN}
12	1.3051	0.0117	0.9468	SsF

3. RESULTS

The PCA of the topostructural indices resulted in the retention of five PCs with eigenvalues greater than or equal to 1.0. These five PCs are presented in Table 1 in terms of their eigenvalues, the amount of data variance incorporated into each of the PCs, and the one topostructural index most-highly correlated with each of the PCs. The topostructural index most-highly correlated with each PC was chosen for use in HiQSAR modeling. Similarly, the PCA on the topochemical indices resulted in 12 PCs with eigenvalues greater than or equal to zero. These PCs and their attendant most-highly correlated topochemical index are presented in Table 2. Table 3 presents definitions for the descriptors used in modeling ARR and D_{37}.

3.1. Results for Modeling ARR. For modeling ARR, all-subsets regression on the set of five topostructural indices resulted in the selection of a three-parameter model. This model used $^4\chi_{PC}$, W, and $DS1_3$, in a poor regression with a regression coefficient (R^2) of 0.3661, a cross-validated regression coefficient (R^2_{cv}) of 0.2213, and a standard error (SE) of 1.27. These three indices were then included in modeling with the topochemical indices, creating a beginning set of 15 topological indices. Again, all-subsets regression was used and resulted in the selection of an eight-parameter model. This model retained W and $DS1_3$ and added the following topochemical indices: $^1\chi^v$, AZV_3, H_{MIN}, SssssC, $SdCH_2$, and SdsCH. This model showed great improvement ($R^2 = 0.8606$, $R^2_{cv} = 0.7637$, SE = 0.63) over the topostructural model. The addition of the six geometrical

Table 3. Definitions of the Descriptors Used in Modeling ARR and D_{37} in *Aspergillus nidulans*

	Topostructural Indices
$^4\chi_{PC}$	path-cluster connectivity index of fourth order
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
$DS1_3$	triplet index from distance matrix, distance sum, and number 1; operation 3
	Topochemical Indices
$^1\chi^v$	valence path connectivity index of first order
$^2\chi^b$	bond path connectivity index of second order
AZV_3	triplet index from adjacency matrix, atomic number, and vertex degree; operation 3
H_{MIN}	minimum H E-state
Q_V	general polarity descriptor
SssssC	electrotopological state index values for carbon with no hydrogen neighbors
$SdCH_2$	electrotopological state index values for terminal double-bonded CH_2
SdsCH	electrotopological state index values for a double-bonded carbon with 1 hydrogen neighbor
SsF	electrotopological state index values for atoms fluorine
	Geometrical Indices
$City_H$	hydrogen-filled city block distance
V_W	van der Waals volume
$^3\kappa\alpha$	kappa alpha index 3
$^3\kappa$	kappa simple index 3
	Quantum Chemical Indices
E_{LUMO}	energy of the lowest unoccupied molecular orbital
E_{LUMO+1}	energy of the second lowest unoccupied molecular orbital
E_{HOMO}	energy of the highest occupied molecular orbital
VEA	vertical electron affinity

indices to this set of eight topological indices result in further improvement in modeling ARR ($R^2 = 0.9014$, $R^2_{cv} = 0.8345$, SE = 0.52). The model selected eliminated all topostructural indices and two topochemical indices and added four geometrical parameters resulting in the following seven-parameter model: H_{MIN}, SssssC, $SdCH_2$, $City_H$, V_W, $^3\kappa\alpha$, and $^3\kappa$. Next, the addition of AM1, STO-3G, 6-31G(d), and 6-311G parameters to the topochemical and geometrical indices did not lead to any improvement in modeling. Modeling with the 6-311G(d) basis set resulted marginal improvement with an eight parameter model, showing a slight increase over the topochemical and geometrical model while using one more index. This model added the energy of the second-lowest unoccupied molecular orbital (E_{LUMO+1}) to the existing hierarchical model. The model statistics are as follows: $R^2 = 0.9065$, $R^2_{cv} = 0.8340$, and SE = 0.51. Finally, the B3lYP/aug-cc-pVTZ parameters were combined with the topochemical and geometrical parameters, resulting in the selection of an eight-parameter model incorporating the following indices: H_{MIN}, SssssC, $SdCH_2$, $City_H$, V_W, $^3\kappa\alpha$, $^3\kappa$, and E_{LUMO}. This model showed some improvement over the topochemical and geometrical model ($R^2 = 0.9120$, $R^2_{cv} = 0.8511$, and SE = 0.50). For the purposes of completeness, models were created using each of the levels of the hierarchy singly. These results, along with the results for the rest of the models, can be found in Table 4.

3.2. Results for Modeling D_{37}. All-subsets regression on the set of five topostructural indices using the SAS procedure REG[79] resulted in the selection of a two-parameter model. This model incorporated W and $DS1_3$, demonstrating a poor regression with a regression coefficient (R^2) of 0.3659, a cross-validated regression coefficient (R^2_{cv}) of 0.2945, and a standard error (SE) of 1.24. These two indices were then included in modeling with the topochemical indices, creating a beginning set of 14 topological indices. Again, all-subsets regression was used and resulted in the selection of a seven-parameter model. This model retained the Weiner index, W, and added the following topochemical indices: AZV_3, $^2\chi^b$, Q_V, $SdCH_2$, H_{MIN}, and SsF. This model was much more satisfactory in terms of R^2 and SE ($R^2 = 0.8791$, $R^2_{cv} = 0.8074$, SE = 0.57). The addition of geometrical indices did not result in any improvement to modeling D_{37}. The model selected during all-subsets regression was identical to the one selected during the previous step in modeling hierarchy. Next, the set of seven topostructural and topochemical parameters was supplemented with the various quantum chemical parameters. The use of AM1 semiempirical descriptors led to the selection of a seven-parameter model showing slight improvement in both R^2 and SE ($R^2 = 0.8840$, $R^2_{cv} = 0.8121$, SE = 0.56). This model replaced the Triplet parameter, AZV_3, with the energy of the second-lowest unoccupied molecular orbital (E_{LUMO+1}). The addition of STO-3G and 6-31G(d) parameters to the topostructural and topochemical indices did not lead to any improvement in modeling. Next, the addition of 6-311G parameters to the seven topostructural and topochemical indices led to some improvement in modeling. A ten-parameter model, adding

Table 4. Summary Results for the HiQSAR Modeling of ARR in *Aspergillus nidulans* for 55 Halogenated Aliphatic Hydrocarbons

model	no. of independent variables	R^2	R^2_{cv}	SE	*F*
TSI only	3	0.3661	0.2213	1.273	9.82
TCI only	8	0.8370	0.7272	0.6795	29.54
3D only	4	0.8326	0.7683	0.6605	62.19
AM1 only	3	0.4285	0.2671	1.208	12.75
STO-3G only	4	0.4240	0.2673	1.225	9.20
6-31G(d) only	4	0.3645	0.1899	1.287	7.17
6-311G only	4	0.4328	0.2790	1.216	9.54
6-311G* only	4	0.5802	0.4337	1.046	17.28
cc-pVTZ only	1	0.4880	0.4504	1.122	50.51
TSI + TCI	8	0.8606	0.7637	0.6285	35.50
TSI + TCI + 3D	7	0.9014	0.8345	0.5228	61.41
TSI + TCI + 3D + AM1	same as previous model				
TSI + TCI + 3D + STO-3G	same as previous model				
TSI + TCI + 3D + 6-31G(d)	same as previous model				
TSI + TCI + 3D + 6-311G	same as previous model				
TSI + TCI + 3D + 6-311G(d)	8	0.9065	0.8340	0.5146	55.77
TSI + TCI + 3D + aug-cc-pVTZ	8	0.9120	0.8511	0.4994	59.59

Table 5. Summary Results for the HiQSAR Modeling of D_{37} in *Aspergillus nidulans* for 55 Halogenated Aliphatic Hydrocarbons

model	no. of independent variables	R^2	R^2_{cv}	SE	F
TSI only	2	0.3659	0.2945	1.243	15.00
TCI only	8	0.8623	0.7749	0.6161	36.00
3D only	8	0.8838	0.6496	0.5660	43.72
AM1 only	3	0.4591	0.3008	1.159	14.43
STO-3G only	3	0.3055	0.1624	1.314	7.48
6-31G(d) only	4	0.4111	0.2458	1.222	8.73
6-311G only	4	0.5140	0.3853	1.110	13.22
6-311G* only	4	0.6318	0.5053	0.9663	21.45
cc-pVTZ only	1	0.5099	0.4787	1.083	55.14
TSI + TCI	7	0.8791	0.8074	0.5710	48.83
TSI + TCI + 3D	same as previous model				
TSI + TCI + AM1	7	0.8840	0.8121	0.5594	51.17
TSI + TCI + STO-3G	7	0.8724	0.7923	0.5866	45.91
TSI + TCI + 6-31G(d)	7	0.8713	0.7926	0.5893	45.45
TSI + TCI + 6-311G	10	0.9015	0.8283	0.5328	40.26
TSI + TCI + 6-311G(d)	8	0.9056	0.8335	0.5100	55.17
TSI + TCI + aug-cc-pVTZ	8	0.9476	0.9236	0.3800	103.98

the energies of the lowest occupied (E_{LUMO}) and second-lowest occupied (E_{LUMO+1}) molecular orbitals and vertical electron affinity (VEA), was selected through all-subsets regression showing a reasonable increase in model performance over the topological model. This model resulted in an $R^2 = 0.9015$, a cross-validated $R^2 = 0.8283$, and SE = 0.53. While this is a noticeable increase in the regression coefficient and the standard error, it is somewhat of a sacrifice with the inclusion of three additional indices. Modeling with the slightly higher 6-311G(d) basis set resulted in a more acceptable eight parameter model, showing a slight increase over the 6-311G model while using two fewer indices. This model removed W, but retained all five of the topochemical indices remaining at this point and added three ab initio parameters: E_{HOMO}, E_{LUMO}, and E_{LUMO+1}. The model statistics are as follows: $R^2 = 0.9056$, $R^2_{cv} = 0.8335$, and SE = 0.51. Finally, the B3LYP/aug-cc-pVTZ parameters were combined with the topostructural and topochemical parameters, resulting in the selection of an eight-parameter model incorporating the following indices: W, AZV_3, ${}^2\chi^b$, Q_V, $SdCH_2$, H_{MIN}, SsF, and E_{LUMO}. This model was by far superior to all of the previous models ($R^2 = 0.9476$, $R^2_{cv} = 0.9236$, and SE = 0.38). For the purposes of completeness, we created models using each of the levels of the hierarchy alone. These results, along with the results for the rest of the models, can be found in Table 5.

4. DISCUSSION

The major objective of this study was to develop high quality QSAR models for the prediction of the toxicity of halocarbons from chemodescriptors calculated directly from their structure. To, this end we have developed HiQSARs for two toxicity endpoints, viz. ARR and D_{37}, for which data were available for the entire set of 55 compounds. Our secondary objective was to examine the contribution of the various levels of molecular descriptors. It is for this reason that we have employed the hierarchical approach to model building. In this way, we examine the ability of each set of descriptors to model the property, and we see how greatly the more complex levels of descriptors contribute to models built from less complex descriptors.

It is clear from the HiQSAR models presented in Tables 4 and 5 that the simplest indices, the TS descriptors, are poor descriptors for modeling these toxicological properties. This is not surprising considering that halocarbons are believed to act through a mechanism dominated by electron transfer to the molecule. The TS indices, being derived from the skeletal graphs of molecules, quantify the connectedness (adjacency and topological distances) of the atoms in the molecular structure. They are completely insensitive to such features as bonding pattern and atom type. The topochemical indices, on the other hand, contribute significantly to the predictive power of the models for both toxicity endpoints. These indices quantify both the topological and chemical aspects of atoms in the molecule, simultaneously characterizing the shape, size, and basic electronic character of molecules. It should be noted that Crebelli et al.[26] found MR to be well correlated with both ARR and D_{37} toxicity values for this set of halocarbons. MR not only is a shape and size descriptor but also encodes information about the electronic nature of the molecule.

The 3-D (geometrical) indices also performed surprisingly well for both properties, though more so in modeling ARR. This should be expected in light of the fact that both the topochemical indices and MR did well in predicting these toxicity endpoints.

In the realm of quantum chemical indices we used both semiempirical (AM1) and ab initio (STO-3G, 6-31G(d), 6-311G, 6-311G(d), and B3LYP/aug-cc-pVTZ) indices. In our model development using the HiQSAR method, we used each of these six classes of quantum chemical indices alone and in the hierarchical modeling with one class of quantum chemical indices being used at a time. As is clear from Table 4, there is little noticeable model improvement for ARR with the addition of quantum chemical parameters of any of the levels included in this study. Even at the high B3LYP/aug-cc-pVTZ level, there is only minor improvement over the existing topochemical and geometrical model. However, as can be seen in Table 5, there is significant improvement in the modeling of D_{37} with the addition of the B3LYP/aug-cc-pVTZ parameters.

It is interesting to note that of the six classes of quantum chemical indices, the high level B3LYP/aug-cc-pVTZ indices were most useful in toxicity prediction either alone or as independent variables in the HiQSAR scheme. This vindicates the belief that the critical step in the toxic mode of action of halocarbons is the attachment of the electron to the halocarbon moiety. The best models after the full hierarchical modeling process yielded the following high quality QSARs:

$$\mathrm{Ln(ARR)} = 39.902 + 0.898(\mathrm{H_{MIN}}) - 0.551(\mathrm{SssssC}) - 0.623(\mathrm{SdCH_2}) + 1.859(\mathrm{City_H}) - 10.726(\mathrm{V_W}) + 7.417({}^3\kappa\alpha) - 8.097({}^3\kappa) + 0.582(\mathrm{E_{LUMO}})$$

$$R^2 = 0.9120,\ R^2_{\mathrm{cv}} = 0.8511,\ n = 55,\ \mathrm{SE} = 0.50,\ F = 59.59$$

$$\mathrm{Ln(D_{37})} = 4.016 - 5.282(\mathrm{W}) + 10.771(\mathrm{AZV_3}) + 2.010({}^2\chi^{\mathrm{b}}) - 0.745(\mathrm{Q_V}) - 0.546(\mathrm{SdCH_2}) + 0.926(\mathrm{H_{MIN}}) + 0.547(\mathrm{SsF}) + 1.329(\mathrm{E_{LUMO}})$$

$$R^2 = 0.9476,\ R^2_{\mathrm{cv}} = 0.9236,\ n = 55,\ \mathrm{SE} = 0.38,\ F = 103.98$$

For comparative purposes, we examined the models initially developed by Crebelli et al.[26] The following equations represent our reanalysis of the Crebelli models (using MR, log*P*, and STO-3G ab initio parameters) to calculate cross-validated regression coefficients:

$$\mathrm{Ln(1/ARR)} = -0.085 + 0.186(\mathrm{MR}) - 10.962(\mathrm{Gap_{LUMO-HOMO}})$$

$$R^2 = 0.6154,\ R^2_{\mathrm{cv}} = 0.5298,\ n = 55,\ \mathrm{SE} = 0.98,\ F = 41.60$$

$$\mathrm{Ln(1/D_{37})} = -0.330 + 0.203(\mathrm{MR}) - 11.249(\mathrm{Gap_{LUMO-HOMO}})$$

$$R^2 = 0.7284,\ R^2_{\mathrm{cv}} = 0.6726,\ n = 55,\ \mathrm{SE} = 0.81,\ F = 69.73$$

The greater success of the B3LYP/aug-cc-pVTZ indices as compared to the STO-3G, 6-31G(d), 6-311G, 6-311G(d), and AM1 parameters in predicting toxicity indicates that more sophisticated or higher level QC indices might better reflect the electronic basis of the toxicity of these halocarbons. To this end, we are currently calculating VEA for a subset of halocarbons at the CCSD(T) level; however, only a small subset will be calculated at this level since is it prohibitively time-consuming to do so for the entire set. We would like to investigate how far the success of the less expensive parameters can be vindicated using the higher level of parameters. Also, we will be doing calculations on solvated molecules as opposed to gas-phase calculations to see whether such involved calculations give better models or shed some light in the toxic modes of action of halocarbons. Such studies are in progress and will be reported subsequently.

ACKNOWLEDGMENT

This study is contribution number 340 from the Center for Water and the Environment of the Natural Resources Research Institute. The authors would like to acknowledge the financial support of this research by the Air Force Office of Scientific Research Grant F49620-02-1-0138 and by Grant/Cooperative Agreement Number 572112 from the Agency for Toxic Substances and Disease Registry (ATSDR). The research at the University of California Davis was supported by the U.S. Department of Energy under Grant No. DEF03-00ER15116. The work done at Lawrence Livermore National Laboratory was performed under the auspices of the U.S. Department of Energy by the University of California, LLNL under contract number W-7405-Eng-48. The contents of the contribution are solely the responsibility of the authors and do not necessarily represent the official views of ATSDR.

Supporting Information Available: Optimized ground state geometries of 55 neutral halocarbons in the Crebelli set (Figure S1), 55 halocarbons and their experimental and estimated toxicity values for ARR and D_{37} (Table S1), ab initio parameters calculated using the MP2/6-31G(d) level of theory (Table S2), ab initio parameters generated within the MP2/6-311G computational scheme (Table S3), ab initio parameters calculated within the MP2/6-311G(d) computational scheme, and ab initio parameters calculated within the b31yp/aug-cc-pVTZ approach. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Benigni, R.; Andreoli, C.; Giuliani, A. QSAR Models for Both Mutagenic Potency and Activity: Application to Nitroarenes and Aromatic Amines. *Environ. Mol. Mutagen.* **1994**, *24*, 208−219.

(2) Mekenyan, O.; Peitchev, D.; Bonchev, D.; Trinajstic, N.; Bangov, I. Modelling the Interaction of Small Organic Molecules with Biomacromolecules. I. Interaction of Substituted Pyridines with Anti-3-Azopyridine Antibody. *Arzneim.-Forsch./Drug Research* **1986**, *36*, 176−183.

(3) Mekenyan, O.; Basak, S. C. In *Graph Theoretic Approaches to Chemical Reactivity*; Bonchev, D., Mekenyan, O., Eds.; Kluwer Academic Publishers: The Netherlands, 1994; pp 221−239.

(4) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Advances in Molecular Similarity, Vol. 2*; Carbo-Dorca, R., Mezey, P. G., Eds.; JAI Press: Stamford, CT, 1998; pp 171−185.

(5) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Quantitative Structure−Activity Relationships in Environmental Sciences VII*; Chen, F., Schuurmann, G., Eds.; SETAC Press: Pensacola, FL, 1998; pp 245−261.

(6) Basak, S. C.; Gute, B. D.; Ghatak, S. Prediction of Complement Inhibitory Activity of Benzamidines Using Topological and Geometrical Parameters. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 255−260.

(7) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Assessment of Mutagenicity of Chemicals from Theoretical Structural Parameters: A Hierarchical Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 117−129.

(8) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999; pp 675−696.

(9) Basak, S. C.; Grunwald, G. D.; Gute, B. D.; Balasubramanian, K.; Opitz, D. Use of Statistical and Neural Net Approaches in Predicting Toxicity of Chemicals. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 885−890.

(10) Basak, S. C.; Gute, B. D.; Grunwald, G. D. In *Discrete Mathematical Chemistry*; Hansen, P., Fowler, P., Zheng, M., Eds.; American Mathematical Society: Providence, RI, 2000; pp 9−24.

(11) Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of Mutagenicity of Aromatic and Heteroaromatic Amines from Structure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 671−678.

(12) Basak, S. C.; Mills, D. Prediction of Mutagenicity Utilizing a Hierarchical Approach. *SAR QSAR Environ. Res.* **2001**, *12*, 481−496.

(13) Basak, S. C.; Mills, D.; Gute, B. D.; Grunwald, G. D.; Balaban, A. T. In *Topology in Chemistry: Discrete Mathematics of Molecules*; Rouvray, D. H., King, R. B., Eds.; Horwood Publishing Ltd.: Chichester, UK, 2001; pp 113−184.

(14) Gute, B. D.; Basak, S. C. Predicting Acute Toxicity of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117−131.

(15) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHS): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1–15.

(16) Roszak, S.; Koski, W. S.; Kaufman, J. J.; Balasubramanian, K. Structures and Electron Attachment Properties of Halomethanes (CX_nY_m, X=H, F.; Y=Cl, Br, I.; N=0,4; M=4-N). *SAR QSAR Environ. Res.* **2001**, *11*, 383–396.

(17) *CRC Handbook of Chemistry and Physics*; CRC Press: Boca Raton, FL, 1996.

(18) Basak, S. C.; Balasubramanian, K. Characterization of Isospectral Graphs Using Graph Invariants and Derived Orthogonal Parameters. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 367–373.

(19) Koski, W. S.; Roszak, S.; Kaufman, J. J.; Balasubramanian, K. Potential Toxicity of Cf_3x Halocarbons. *In Vitro Toxicol.* **1998**, *10*, 455–457.

(20) Roszak, S.; Koski, W. S.; Kaufman, J. J.; Balasubramanian, K. Structure and Energetics of CF_3Cl^-, CF_3Br^-, CF_3I^- Radical Ions. *J. Chem. Phys.* **1997**, *106*, 7709–7713.

(21) Cash, G. G. personal communication, 2001.

(22) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure–Activity Relationships (SAR) under TSCA, Section 5. *Environ. Health Perspect.* **1990**, *87*, 183–197.

(23) Fishbein, L. In *Environmental Carcinogens – Selected Methods of Analysis*; Fishbein, L., O'Neill, I. K., Eds.; International Agency for Research on Cancer: Lyon, 1985; Vol. 7, pp 47–67.

(24) Pearson, C. R. In *Handbook of Environmental Chemistry*; Hutzinger, O., Ed.; Springer-Verlag: New York, 1982; Vol. 3, pp 69–88.

(25) Stephenson, M. E. An Approach to the Identification of Organic Compounds Hazardous to the Environment and Human Health. *Ecotoxicol. Environ. Safety* **1977**, *1*, 407–425.

(26) Crebelli, R.; Andreoli, C.; Carere, A.; Conti, L.; Crochi, B.; Cotta-Ramusino, M.; Benigni, R. Toxicology of Halogenated Aliphatic Hydrocarbons: Structural and Molecular Determinants for the Disturbance of Chromosome Segregation and the Induction of Lipid Peroxidation. *Chem.-Biol. Interact.* **1995**, *98*, 113–129.

(27) Basak, S. C.; Grunwald, G. D. Predicting Mutagenicity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* **1995**, *31*, 2529–2546.

(28) Basak, S. C.; Gute, B. D.; Grunwald, G. D. A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol–Water Partition Coefficient. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054–1060.

(29) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651–655.

(30) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997; pp 73–116.

(31) Basak, S. C.; Mills, D. Quantitative Structure–Property Relationships (QSPRs) for the Estimation of Vapor Pressure: A Hierarchical Approach Using Mathematical Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 692–701.

(32) Crebelli, R.; Andreoli, C.; Carere, A.; Conti, G.; Conti, L.; Cotta-Ramusino, M.; Benigni, R. The Induction of Mitotic Chromosome Malsegregation in *Aspergillus Nidulans.* Quantitative Structure Activity Relationship (QSAR) Analysis with Chlorinated Aliphatic Hydrocarbons. *Mutat. Res.* **1992**, *266*, 117–134.

(33) Benigni, R.; Andreoli, C.; Conti, L.; Tafani, P.; Cotta-Ramusino, M.; Carere, A.; Crebelli, R. Quantitative Structure–Activity Relationship Models Correctly Predict the Toxic and Aneuploidizing Properties of Halogenated Methanes in *Aspergillus Nidulans*. *Mutagenesis* **1993**, *8*, 301–305.

(34) *Polly*, copyright of the University of Minnesota, 1988.

(35) *Molconn-Z*, v 3.50, Hall Associates Consulting Quincy, MA, 2000.

(36) Filip, P. A.; Balaban, T. S.; Balaban, A. T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlational Ability. *J. Math. Chem.* **1987**, *1*, 61–83.

(37) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(38) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(39) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, U.K., 1986.

(40) Bonchev, D.; Trinajstic, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.

(41) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of Isomeric Structures Using Information Theoretic Topological Indices. *J. Comput. Chem.* **1984**, *5*, 581–588.

(42) Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the Structure–Function Relationship of Pharmacological and Toxicological Agents Using Information Theory, University of Missouri–Rolla: Rolla, MO, 1980.

(43) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis: A Quantitative Structure–Activity Relationship (QSAR) Study of Alcohols Using Complementary Information Content (CIC). *Arzneim.-Forsch./Drug Res.* **1983**, *33*, 501–503.

(44) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Math. Modelling Sci. Technol.*; Avula, X. J. R., Kalman, R. E., Liapis, A. I., Rodin, E. Y., Eds.; Pergamon Press: 1984; pp 745–750.

(45) Balaban, A. T. Highly Discriminating Distance-Based Topological Indices. *Chem. Phys. Lett.* **1982**, *89*, 399–404.

(46) Balaban, A. T. Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199–206.

(47) Balaban, A. T. Chemical Graphs. Part 48. Topological Index J for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH)* **1986**, *21*, 115–122.

(48) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.

(49) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math.* **1988**, *19*, 17–44.

(50) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(51) Bogdanov, B.; Nikolic, S.; Trinajstic, N. On the Three-Dimensional Wiener Number. *J Math Chem.* **1989**, *3*, 299–309.

(52) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109–116.

(53) Kier, L. B.; Hall, L. H. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach Science Publishers: Amsterdam, 1999; pp 455–489.

(54) Tripos Associates, Inc., St. Louis, MO, 1997.

(55) Tripos Associates, Inc., St. Louis, MO, 1997.

(56) Stewart, J. J. P. MOPAC Version 6.00, QCPE #455, Frank J. Seiler Research Laboratory, U.S. Air Force Academy, CO, 1990.

(57) Møller, C.; Plesset, M. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1943**, *46*, 618–622.

(58) Hehre, W. J.; Stewart, R. F.; Pople, J. A. *J. Chem. Phys.* **1969**, *51*, 2657.

(59) Collins, J. B.; Schleyer, P. v. R.; Binkley, J. S.; Pople, J. A. Self-Consistent Molecular Orbital Methods. 17. Geometries and Binding Energies of Second-Row Molecules. A Comparison of Three Basis Sets. *J. Chem. Phys.* **1976**, *64*, 5142.

(60) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724.

(61) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.

(62) Hariharan, P. C.; Pople, J. A. *Mol. Phys.* **1974**, *27*, 209.

(63) Gordon, M. S. *Chem. Phys. Lett.* **1980**, *76*, 163.

(64) Hariharan, P. C.; Pople, J. A. *Theo. Chim. Acta* **1973**, *28*, 213.

(65) Binning, J. R., Jr.; Curtiss, L. A. *J. Comput. Chem.* **1990**, *11*, 1206.

(66) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639.

(67) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

(68) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.

(69) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

(70) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(71) Peterson, K. A.; Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 7410.

(72) Wilson, A.; van Mourik, T.; Dunning, T. H., Jr. *J. Mol. Struct. (THEOCHEM)* **1997**, *388*, 339.

(73) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(74) Becke, A. D. Density Functional Theory Thermochemistry. 3. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(75) Vosko, S. H.; Wilk, L.; Nusiar, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: A Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(76) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy into a Functional Theory of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785–789.

(77) Brown, S. T.; Rienstra-Kiracofe, J. C.; Schaefer, H. F. *J. Phys. Chem. A* **1999**, *103*, 4065.

(78) *Gaussian 98 (Revision A.11.2)*; Gaussian, Inc.: Pittsburgh, PA, 1998.

(79) *Sas/Stat User's Guide, Release 6.03*; SAS Institute Inc.: Cary, NC, 1988.

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

CI020054N