Prediction of Hydroxyl Radical Rate Constants from Molecular Structure

Gregory A. Bakken and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, 152 Davey Laboratory, University Park, Pennsylvania 16802

Received April 15, 1999

Quantitative structure−property relationships are developed using multiple linear regression and computational neural networks (CNNs). Structure-based descriptors are used to numerically encode molecular features that can be used to form models describing reaction rates with hydroxyl radicals. For a set of 57 unsaturated hydrocarbons, a 5−2−1 CNN was developed that produced a root-mean-square (rms) error of 0.0638 log units for the training set and 0.0657 log units for an external prediction set. The residual sum of squares for all 57 compounds was 0.234 log units, which compares very favorably with existing methodologies. Additionally, a 10−7−1 CNN was built to predict hydroxyl radical rate constants for a diverse set of 312 compounds. The training set rms error was 0.229 log units, and the rms error for the external prediction set was 0.254 log units. This model demonstrates the ability to provide accurate predictions over a wide range of functionalities.

INTRODUCTION

The reactions of organic pollutants in the atmosphere are of great concern from an ecological standpoint. Risk assessment measurements, such as degradation pathways and atmospheric lifetimes, provide information regarding the fate of these compounds. Experimental determination of risk assessment measures is often a time consuming, expensive process. Therefore, estimation techniques that accurately model risk assessment parameters are very desirable.

Hydroxyl radicals provide the major degradation pathway for most organic compounds in the atmosphere.[1] Therefore, much experimental work has been done to determine rate constants for reaction of hydroxyl radicals with organic substrates.[1−3] To reduce analysis time and cost associated with experimentally measuring rate constants, several methods have been developed to form models to estimate these values.[4−17] Estimation methods for hydroxyl radical reaction rates have been developed using such things as bond dissociation energy,[4−7] ionization potentials,[8,9] fragment contribution methods,[10−13] and molecular orbital calculations.[14,15] A recent paper has described and evaluated some of the existing methods.[16] All of these methods have been at least moderately successful in forming prediction models, but little work has been done to examine model validation over a wide range of compounds.

Hydroxyl radical reactions are typically grouped into four reaction types: (1) hydrogen atom abstraction, (2) addition to double and triple bonds, (3) addition to aromatic rings, and (4) reactions with nitrogen, sulfur, and phosphorus.[1,13] When prediction models are developed, each of these four processes can be dealt with separately, or all four reaction types can be grouped together. The majority of work to date has focused on examining each reaction type individually.[12−15]

Recently, a quantitative structure−property relationship (QSPR) was developed using partial least squares with empirical and calculated quantum chemical descriptors.[17] The model was developed for a set of 57 unsaturated hydrocarbons, so all reactions were via hydroxyl radical addition to a multiple bond. All compounds were used in model formation. Randomization of the dependent variable was used to verify a true correlation was being modeled. A comparison was made to Klamt's QSPR model[14] based on quantum chemical descriptors and to Atkinson's fragment contribution method[12] using the 57 unsaturated hydrocarbons. Atkinson's fragment contribution method produced the lowest residual sum of squares (RSS) error for the compounds examined. However, this result may be somewhat biased since many of the 57 available compounds were used to calculate the fragment contribution parameters. A drawback of these three approaches is that unique models are formed for specific reaction classes instead of forming a single model for all compounds.

This paper presents a QSPR developed for the 57 unsaturated compounds mentioned above. Hydroxyl radical rate constants (k_{OH}) are related to molecular structure via structure-based descriptors. Selected descriptor subsets are used with multiple linear regression (MLR) and computational neural networks (CNNs) for model formation. The methodology employed has been used in forming QSPRs for prediction of a wide range of properties, including aqueous solubility,[18] human intestinal absorption,[19] vapor pressure,[20] and liquid crystal clearing temperature.[21] Additionally, a recent study demonstrated a QSPR for methyl radical addition rate constants that provided prediction on the order of experimental error.[22] Results obtained in the present study will be compared to the results obtained in ref 17 using PLS, Klamt's method, and Atkinson's method.

In addition to the set of 57 unsaturated hydrocarbons, a QSPR is developed to predict k_{OH} for a diverse set of 312 compounds. All four of the hydroxyl radical reaction classes discussed are present in this set. This set includes alkenes, alcohols, halogens, nitrates, amines, aromatics, and other functionalities. This set was examined to explore the possibility of developing one QSPR for all organic compounds

10.1021/ci990042a CCC: $18.00
Published on Web 09/18/1999

falling into the four reaction classes described instead of forming a unique model for each compound class encountered. This will allow some examination of the predictive ability of the model over a wide range of compounds. Additionally, the model developed is used to predict rate constants for the subset of 57 unsaturated hydrocarbons. This allows quantitative comparison of predictions using one model for all compounds and developing unique models for structurally similar compounds.

EXPERIMENTAL SECTION

The k_{OH} values used in this study were obtained from the literature.[12,14,17] Data selected were for reactions run at 298 K. Therefore, predictions obtained using these compounds are for rates at this temperature. Compounds examined in this study are presented in Tables 1 and 2, along with experimentally determined k_{OH} values. Presented values are the negative logarithm of the actual rates in units of cm^3 $molecule^{-1}$ s^{-1}.

Table 1 presents the 57 compounds selected to comprise data set 1. These 57 compounds are unsaturated hydrocarbons, and they were selected for study so comparison could be made to existing methodology.[17] The 57 compounds were randomly divided to form a 52-member training set and a 5-member external prediction set. Minimization of the root-mean-square (rms) error of the training set compounds was used to direct the search for a descriptor subset forming a linear model. Once a model was developed, the external prediction set was used to validate the model and examine predictive ability.

Nonlinear models were generated using CNNs. Five compounds were randomly selected and removed from the training set and placed in a cross-validation set, leaving 47 training set compounds, 5 cross-validation set compounds, and 5 prediction set compounds. The cross-validation set was necessary to avoid overtraining the CNN. The training process, and the role of the cross-validation set, will be discussed more fully in Results and Discussion. The prediction set was again used to demonstrate the ability of the model to generalize.

All 312 compounds in Table 2 were used in generating a QSPR (data set 2). This larger set was selected to examine the diversity that could be modeled with a QSPR. Additionally, comparison of predictions of the 57 compounds in data set 1 with models built with data set 1 and models built with data set 2 illustrates the relationship between accurate predictions and model diversity. For linear model formation, the training set contained 281 compounds and the external prediction set contained 31 compounds. Nonlinear CNN models were generated with 250 members in the training set, 31 members in the cross-validation set, and 31 members in the prediction set.

QSPRs were developed using the Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software system.[23,24] Feature selection routines (genetic algorithm[25] and simulated annealing[26−28]) and CNN routines[29] were written in-house. All computations were preformed on a DEC 3000 AXP Model 500 workstation. Methods used to develop QSPRs can be broken into five basic steps: (1) structure entry and modeling, (2) descriptor generation, (3) objective feature selection, (4) linear model formation and validation, and (5) nonlinear model formation and validation.

Structure Entry and Modeling. Two-dimensional sketches of compounds were obtained using HyperChem (Hypercube, Inc., Waterloo, ON), and initial coordinates for three-dimensional models were obtained. Information was stored in connection tables containing atom types, bond angles, and bonding information. Accurate three-dimensional models are required for calculation of certain descriptors, so the initial coordinates determined using HyperChem were passed to the semiempirical molecular orbital program MOPAC[30] for refinement. A PM3 Hamiltonian[31] was selected for geometry optimization.

Descriptor Generation. Formation of a successful QSPR requires descriptors that accurately describe the property of interest. Therefore, it is paramount that an information-rich pool of descriptors be available. In this study, a total of 196 structure-based descriptors were calculated for data set 1 and 208 for data set 2. Calculated descriptors can be divided into four classes: topological, geometric, electronic, and hybrid or combination descriptors. Of the 196 descriptors computed for data set 1, 113 were topological, 27 were geometric, 8 were electronic, and 48 were combination descriptors. For data set 2, 123 were topological, 29 were geometric, 8 were electronic, and 48 were combination.

Topological descriptors[32−34] are calculated on the basis of a simple two-dimensional sketch of the molecule; i.e., no geometry optimization is required. Descriptors calculated included molecular connectivity, fragment, and κ indices. Molecular connectivity values encoded information pertaining to the size and degree of branching. Fragment descriptors provided counts of atom types, counts of bonds, counts of rings, etc. κ indices provided information about molecular shape using only a two-dimensional molecular sketch.

Geometric descriptors[35,36] require the coordinates for energy minimized three-dimensional structures. Calculated geometric descriptors included moments of inertia, gravitational index, and principal axes of the molecule. Additionally, solvent-accessible surface areas and volumes were encoded.

The third category considered is electronic descriptors.[37−39] These descriptors encoded the electronic environment of the molecule. Examples of electronic descriptors included partial atomic charges, energy of highest occupied molecular orbital, and energy of lowest unoccupied molecular orbital.

Hybrid or combination descriptors represent the final class of descriptors used. Information is taken from at least two of the classes previously described to compute the combination descriptors. For example, information from the electronic descriptor for partial atomic charges was combined with information from the geometric descriptor for solvent accessible surface area to form charged partial surface area descriptors.[40] Additional hybrid descriptors were used to encode information about hydrogen bonding.

Objective Feature Selection. Feature selection can be classified as objective or subjective. Objective feature selection involves eliminating descriptors solely on the basis of descriptor values.

Subjective feature selection, which will be addressed in the sections on model formation, utilizes dependent variable information when selecting descriptors. Before subjective feature selection is carried out, objective means should be used to reduce the descriptor pool to ensure the ratio of descriptors to compounds is less than or equal to 0.6. This will work to prevent chance correlations from being modeled.

Table 1. Compounds Comprising Data Set 1

no.	compound	$-\log(k_{OH})$[a]	PLS[b]	Klamt[b]	Atkinson[b]	type 1[c]	type 2[c]	type 3[c]
1	α-terpinene	9.44	9.47	9.63	9.63	9.61	9.57	9.65
2	α-phellandrene	9.50	9.62	9.73	9.67	9.63	9.60	9.69
3	*trans*-*β*-ocimene	9.60	9.68	9.65	9.59	9.73	9.73	9.56
4	terpinolene	9.65	9.67	9.70	9.55	9.35[p]	9.68[p]	9.72[p]
5	myrcene	9.67	9.90	9.71	9.68	9.58	9.72	9.67
6	2,5-dimethyl-2,4-hexadiene	9.68	9.60	9.64	9.60	9.67	9.66	9.67
7	*γ*-terpinene	9.75	9.73	9.75	9.68	9.60	9.57	9.71
8	D-limonene	9.77	9.88	9.84	9.78	9.67	9.65	9.77
9	*β*-phellandrene	9.78	9.77	9.84	9.76	9.74	9.66[cv]	9.80[cv]
10	1,3-cyclohexadiene	9.79	9.83	9.86	9.82	9.90	9.93	9.85
11	*trans*,*trans*-2,4-hexadiene	9.87	9.88	9.87	9.77	9.90	9.87	9.84
12	*trans*-4-methyl-1,3-pentadiene	9.88	9.82	9.87	9.76	9.99	9.86	9.86
13	2,3-dimethyl-1,3-butadiene	9.91	9.98	9.87	9.90	9.86[p]	9.96[p]	9.94[p]
14	2,5-dimethyl-1,5-hexadiene	9.92	10.01	9.98	9.97	9.86	9.96	9.86
15	bicyclo[2.2.1]-2,5-heptadiene	9.92	10.17	9.94	9.93	9.87	9.92	9.88
16	*trans*-1,3-hexadiene	9.95	9.97	9.98	9.87	10.08	9.99	9.94
17	*trans*-1,3,5-hexatriene	9.96	9.88	9.95	9.80	10.00	9.97	9.96
18	*cis*-1,3,5-hexatriene	9.96	9.90	9.95	9.79	10.00	9.97	9.98
19	2,3-dimethyl-2-butene	9.96	9.84	9.96	9.79	9.86	9.96	9.91
20	1,3-pentadiene	10.00	10.00	9.98	9.88	10.10	10.00	9.98
21	2-methyl-1,3-butadiene	10.00	10.09	9.98	9.99	10.15	10.02	10.05
22	1,4-cyclohexadiene	10.00	10.01	9.94	9.88	9.89	9.93	10.00
23	1,3,5-cycloheptatriene	10.01	9.79	9.79	9.73	9.71	9.95	9.99
24	2-methyl-1,5-hexadiene	10.02	10.17	10.10	10.06	10.06	10.01	10.05
25	*trans*-1,4-hexadiene	10.04	10.22	10.04	10.03	10.04	9.97[cv]	10.10[cv]
26	2-methyl-2-pentene	10.05	9.97	10.05	9.97	10.10[p]	9.99[p]	10.01[p]
27	Δ3-carene	10.06	9.90	10.05	9.93	10.10	10.15	10.08
28	2-methyl-2-butene	10.06	10.02	10.06	9.99	10.12	10.00	10.04
29	*β*-pinene	10.10	10.09	10.25	10.11	10.22	10.21	10.23
30	cycloheptene	10.13	10.12	10.20	10.08	10.16	10.10	10.21
31	*trans*-4-octene	10.16	10.13	10.17	10.15	10.16	10.13[cv]	10.12[cv]
32	*trans*-2-heptene	10.17	10.16	10.17	10.16	10.20	10.16	10.23
33	cyclohexene	10.17	10.16	10.21	10.10	10.18	10.11	10.16
34	cyclopentene	10.17	10.19	10.23	10.12	10.20	10.13	10.19
35	*trans*-2-pentene	10.18	10.26	10.19	10.18	10.22	10.18	10.22
36	1,3-butadiene	10.18	10.26	10.18	10.07	10.30	10.18	10.16
37	*cis*-2-pentene	10.18	10.20	10.24	10.15	10.22	10.18	10.19
38	*trans*-2-butene	10.19	10.30	10.19	10.18	10.23	10.21	10.23
39	2-methyl-1-pentene	10.20	10.20	10.27	10.26	10.27	10.27[cv]	10.30[cv]
40	1,5-hexadiene	10.21	10.40	10.26	10.17	10.22	10.14	10.18
41	2-methyl-1-butene	10.22	10.25	10.28	10.29	10.27	10.27[cv]	10.30[cv]
42	*trans*-4-methyl-2-pentene	10.22	10.22	10.18	10.17	10.08	10.18	10.21
43	3-methyl-1,2-butadiene	10.25	10.24	10.24	10.25	10.27	10.27	10.14
44	*cis*-2-butene	10.25	10.25	10.25	10.18	10.23	10.21	10.22
45	α-pinene	10.27	9.89	10.05	9.87	10.10	10.16	10.12
46	1,4-pentadiene	10.27	10.44	10.27	10.20	10.25[p]	10.16[p]	10.23[p]
47	camphene	10.27	10.17	10.23	10.27	10.20	10.18	10.16
48	2-methylpropene	10.29	10.28	10.29	10.27	10.27[p]	10.27[p]	10.26[p]
49	1-heptene	10.39	10.36	10.50	10.42	10.43	10.47	10.46
50	1-hexene	10.43	10.40	10.52	10.43	10.43	10.48	10.45
51	1,2-pentadiene	10.45	10.40	10.50	10.37	10.57	10.53	10.48
52	3-methyl-1-butene	10.50	10.47	10.55	10.46	10.31	10.54	10.52
53	1-pentene	10.50	10.45	10.54	10.45	10.43	10.48	10.44
54	1-butene	10.50	10.50	10.56	10.47	10.44	10.49	10.45
55	3,3-dimethyl-1-butene	10.55	10.46	10.57	10.53	10.52	10.56	10.46
56	1,2-butadiene	10.58	10.45	10.46	10.39	10.57	10.56	10.55
57	propadiene	11.01	10.72	10.77	10.58	10.69	10.84	11.03
	RSS		0.710	0.781	0.382	0.676	0.281	0.234

[a] Units of cm^3 $molecule^{-1}$ s^{-1}. [b] Data taken from ref 17. [c] Superscript abbreviations: cv = cross-validation set members; p = prediction set members.

Objective feature selection was performed using only training set compounds (52 for data set 1 and 281 for data set 2). Any descriptor with identical values for more than 90% of the observations in the training set was eliminated. Obviously, such descriptors carried no useful information. Pairwise correlations were used to reduce redundant information. One of any two descriptors with $r > 0.93$ was removed from the pool. This left 59 descriptors for data set 1 and 109 descriptors for data set 2. The reduced pool for data set 2 is acceptable, but before subjective feature selection, more descriptors must be removed from the data set 1 pool to lower the ratio of descriptors to compounds to below 0.6.

To further reduce the pool for data set 1, orthogonality of the descriptors remaining in the pool was examined using vector space descriptor analysis.[41] A single descriptor was

Table 2. Compounds Comprising Data Set 2

no.	compound	$-\log(k_{OH})^a$	type 1[b]	type 2[b]	type 3[b]	ref	no.	compound	$-\log(k_{OH})^a$	type 1[b]	type 2[b]	type 3[b]	ref
1	ethane	12.57	12.30	12.35[cv]	12.84[cv]	14	70	dihydroxydiethylether	10.52	10.68	10.54	10.69	12
2	*n*-butane	11.60	11.55	11.45[cv]	11.58[cv]	14	71	2-ethoxyethanol	10.92	10.77[p]	10.58[p]	10.86[p]	12
3	2,2-dimethylpropane	12.07	12.50	11.84[cv]	12.43[cv]	14	72	2-butoxyethanol	10.85	10.57	10.44	10.82	12
4	*n*-pentane	11.40	11.46	11.34	11.38	14	73	dimethyl ether	11.53	11.33[p]	11.27[p]	11.29[p]	12,14
5	2-methylpropane	11.63	11.92	11.72	11.64	14	74	diethyl ether	10.88	10.90	11.04	10.85	14
6	*n*-propane	11.94	11.99	11.83	11.95	14	75	dipropyl ether	10.76	10.72	10.97	10.69	14
7	cyclopentane	11.29	11.53	11.21	11.23	14	76	*tert*-butyl methyl ether	11.55	11.14	11.51	11.39	14
8	2-methylbutane	11.41	11.44	11.32	11.39	14	77	tetrahydrofuran	10.79	11.14	10.79	10.81	14
9	2-methylpentane	11.25	11.31	11.24	11.26	14	78	ethylene oxide	13.15	11.96	12.72	12.85	12
10	3-methylpentane	11.24	11.35	11.19	11.27	14	79	1,3,5-trioxane	11.10	11.38	10.97	10.84	12
11	2,2-dimethylbutane	11.63	11.68	11.70	11.79	14	80	propylene oxide	12.28	11.77	12.30	12.29	12
12	2,3-dimethylbutane	11.21	11.24	11.09	11.12	14	81	1,2-epoxybutane	11.68	11.65	11.91[cv]	11.76[cv]	12
13	2,4-dimethylpentane	11.29	11.21	11.17	11.17	14	82	2-chloroethanol	11.85	11.89	12.21	12.26	12
14	2,2,3-trimethylbutane	11.37	11.54	11.42	11.45	14	83	1-chloro-2,3-epoxy-propane	12.36	12.21[p]	12.86[p]	12.57[p]	12
15	2,2,4-trimethylpentane	11.43	11.30	11.37[cv]	11.34[cv]	14							
16	2,2,3,3-tetramethylbutane	11.97	11.95	11.99	11.72	14	84	methyl acetate	12.77	11.75	12.26	12.60	12
17	cyclohexane	11.13	11.35	11.07	11.05	14	85	ethyl acetate	11.77	11.61	11.90	12.18	12
18	methylcyclohexane	10.98	11.11	10.91	11.08	14	86	propyl acetate	11.60	11.46	11.63	11.77	12
19	*n*-octane	11.06	11.07	11.14	11.07	14	87	isopropyl acetate	11.51	11.44	11.56	11.37	12
20	*n*-decane	10.94	10.83	11.09	10.95	14	88	butyl acetate	11.37	11.29	11.35	11.47	12
21	cycloheptane	10.91	11.17	10.97	11.02	14	89	2-ethoxyethyl acetate	10.89	11.21	11.48	11.25	12
22	bicyclo[2.2.1]heptane	11.26	11.07[p]	10.82[p]	11.07[p]	14	90	isobutyl acetate	11.27	11.30	11.36	11.30	12
23	bicyclo[2.2.2]octane	10.83	10.85	10.70	11.13	14	91	methyl propionate	12.57	11.61	12.21	11.97	12
24	fluoromethane	13.77	12.82	13.39	13.84	14	92	ethyl propionate	11.78	11.49	11.94	11.63	12
25	chloromethane	13.36	12.43	12.78	13.13	14	93	trichloroacetaldehyde	11.71	12.63	12.04	11.77	12
26	difluoromethane	13.96	13.17	13.91	13.91	14	94	acetyl chloride	13.17	12.59	13.29	13.15	12
27	dichloromethane	12.85	12.90[p]	13.29[p]	12.67[p]	14	95	ethyl nitrate	12.31	12.30	12.40	12.51	12
28	chlorofluoromethane	13.36	12.97[p]	13.46[p]	13.22[p]	14	96	*n*-propyl nitrate	12.14	12.14	12.16	12.19	12
29	trifluoromethane	15.70	14.42	14.85	15.44	14	97	isopropyl nitrate	12.74	12.02	11.99	12.26	12
30	chlorodifluoromethane	14.30	13.84	14.24	14.43	14	98	*n*-butyl nitrate	11.86	11.99	11.90	12.06	12
31	dichlorofluoromethane	13.52	13.66	14.01	13.49	14	99	2-butyl nitrate	12.17	12.00	11.90	12.01	12
32	trichloromethane	12.99	13.51	13.81[cv]	12.84[cv]	14	100	2-pentyl nitrate	11.74	11.85	11.77	11.88	12
33	fluoroethane	12.64	12.49	12.84	12.63	14	101	3-pentyl nitrate	11.96	11.91[p]	11.78[p]	11.95[p]	12
34	chloroethane	12.41	12.32	12.54	12.57	14	102	3-methyl-2-butyl nitrate	11.76	11.80	11.73	11.84	12
35	1,1-difluoroethane	13.47	13.13	13.83	13.44	14	103	2,2-dimethylpropyl nitrate	12.07	12.12	12.09	12.06	12
36	1,2-difluoroethane	12.96	12.75	13.16	12.90	14	104	2-hexyl nitrate	11.50	11.56	11.54	11.65	12
37	1,1-dichloroethane	12.59	12.88	13.22	12.67	14	105	3-hexyl nitrate	11.58	11.74	11.66[cv]	11.76[cv]	12
38	1,2-dichloroethane	12.66	12.57	12.65	12.54	14	106	cyclohexyl nitrate	11.48	11.87	11.68	11.29	12
39	1,1,1-trifluoroethane	14.70	14.35	14.82[cv]	14.73[cv]	14	107	2-methyl-2-pentyl nitrate	11.77	11.79	11.86	11.71	12
40	1,1,2,2-tetrafluoroethane	13.30	13.65	13.67	13.46	14	108	3-methyl-2-pentyl nitrate	11.52	11.52	11.60	11.54	12
41	1-chloro-1,1-difluoroethane	14.40	13.87	14.22	14.46	14	109	3-heptyl nitrate	11.44	11.47	11.50[cv]	11.55[cv]	12
42	1,1,1-trichloroethane	13.92	13.71	13.96	13.88	14	110	3-octyl nitrate	11.42	11.16	11.35	11.40	12
43	1,1,2-trichloroethane	12.50	12.92	12.86	12.36	14	111	tetrahydropyran	10.86	10.97	10.60	10.60	14
44	1,1,1,2-tetrafluoroethane	14.05	14.00	13.97	14.19	14	112	1,3-dioxane	11.04	11.25	11.39	10.99	14
45	1-chloro-2,2,2-trifluoroethane	13.80	13.52	13.41	13.71	14	113	1,4-dioxane	10.96	10.86	10.77	10.51	14
46	1,2-dichloro-1,1-difluoroethane	13.64	13.47	13.35	13.66	14	114	1,1-dimethoxyethane	11.05	11.06[p]	11.11[p]	11.20[p]	14
47	pentafluoroethane	14.40	14.43	14.01[cv]	14.30[cv]	14	115	2,2-dimethoxypropane	11.41	11.23	11.39	11.37	14
48	1-chloro-1,2,2,2-tetrafluoroethane	14.00	14.14	13.82	13.88	14	116	1,2-dimethoxypropane	10.84	10.71	10.59[cv]	10.75[cv]	14
							117	methylamine	10.66	10.73	10.31	10.43	12
49	1,1-dichloro-2,2,2-trifluoroethane	13.48	13.97	13.78	13.51	14	118	ethylamine	10.56	10.71[p]	10.29[p]	10.42[p]	12
							119	dimethylamine	10.18	10.69	10.59	10.28	12
50	4-methyloctane	11.01	10.98	11.08	11.02	14	120	trimethylamine	10.22	10.56[p]	10.51[p]	10.28[p]	12
51	2,3,5-trimethylhexane	11.10	10.71	10.89	10.95	14	121	diethylhydroxylamine	10.00	10.34	10.28	10.47	12
52	cyclooctane	10.86	10.99	10.93	10.87	14	122	diethylaminoethanol	10.10	9.98	10.06	9.85	12
53	1,1,3-trimethylcyclohexane	11.06	11.11	11.05	11.00	14	123	2-amino-2-methyl-1-propanol	10.55	10.36	10.70	10.62	12
54	bicyclo[3.3.0]octane	10.96	11.15	10.92[cv]	10.85[cv]	14							
55	tricyclo[3.3.1.1]decane	10.65	10.56	10.48[cv]	10.85[cv]	14	124	*N*-nitrosodimethylamine	11.60	10.96	10.97	11.40	12
56	1-chloropropane	12.10	11.98	12.01	12.43	14	125	*N*-nitrodimethylamine	11.42	11.46	11.66	11.43	12
57	2-chloropropane	12.37	12.30	12.46	12.52	14	126	hydrazine	10.21	10.78	10.24	10.23	12
58	2,2-dichloropropane	13.10	13.13	13.15	12.91	14	127	monomethylhydrazine	10.19	10.57	10.30	10.31	12
59	1,2-dichloropropane	12.36	12.33	12.21	12.27	14	128	methanethiol	10.48	10.71	10.69[cv]	10.80[cv]	12
60	1,2,3-trichloropropane	12.37	12.76	12.66	12.37	14	129	ethanethiol	10.33	10.71	10.65	10.67	12
61	1,3-dichloropropane	12.22	12.48	12.46	12.36	14	130	propanethiol	10.34	10.44	10.54	10.56	12
62	methanol	12.05	11.61	11.83	12.12	12	131	2-propanethiol	10.37	10.68	10.56	10.57	12
63	ethanol	11.54	11.45	11.49	11.48	12	132	butanethiol	10.36	10.38	10.39	10.44	12
64	propanol	11.28	11.17[p]	11.59[p]	11.31[p]	12	133	2-methyl-1-propanethiol	10.38	10.36	10.38[cv]	10.42[cv]	12
65	2-propanol	11.21	11.49	11.58	11.39	12	134	2-butanethiol	10.40	10.31	10.31	10.40	12
66	2-methyl-2-propanol	11.96	11.84	12.13	11.98	12	135	*tert*-butanethiol	10.45	10.99	10.35	10.41	12
67	butanol	11.14	11.12	11.42	11.20	12	136	3-methyl-1-butanethiol	10.28	10.23	10.29	10.26	12
68	1,2-dihydroxyethane	11.11	11.07	11.02	11.01	12	137	dimethyl sulfide	11.37	10.46	10.75	10.86	12
69	1,2-dihydroxypropane	10.92	11.05	11.02	11.23	12	138	ethyl methyl sulfide	11.07	10.17	10.60	10.70	12

Table 2. (Continued)

no.	compound	− log (k_{OH})[a]	type 1[b]	type 2[b]	type 3[b]	ref	no.	compound	− log (k_{OH})[a]	type 1[b]	type 2[b]	type 3[b]	ref
139	diethyl sulfide	10.81	10.13[p]	10.48[p]	10.53[p]	12	204	2-methyl-1-pentene	10.20	10.22	10.08	10.32	14,17
140	tetrahydrothiophene	10.72	10.28	10.40[cv]	10.54[cv]	12	205	2-methyl-2-pentene	10.05	10.11	10.00[cv]	10.02[cv]	14,17
141	dimethyldisulfide	9.69	10.32	10.26	10.03	12	206	*trans*-4-methyl-2-pentene	10.22	10.37	10.18	10.12	14,17
142	fluoroethene	11.25	11.55	11.12	10.97	12	207	2,3-dimethyl-1,3-butadiene	9.91	9.80[p]	9.83[p]	10.18[p]	14,17
143	chloroethene	11.18	11.60	11.16	10.98	12,14	208	*cis*-1,3,5-hexatriene	9.96	9.67	9.96	9.86	14,17
144	bromoethene	11.17	11.56	11.07	10.98	12,14	209	*trans*-β-ocimene	9.60	9.09	9.56	9.71	17
145	1,1-dichloroethene	10.94	11.91	11.48	10.99	14	210	terpinolene	9.65	9.48	9.80	9.68	17
146	*cis*-1,2-dichloroethene	11.62	11.54	11.14	11.72	12,14	211	myrcene	9.67	9.23	9.58	9.68	17
147	*trans*-1,2-dichloroethene	11.74	11.56	11.17	11.71	12,14	212	β-phellandrene	9.77	9.41	9.56	9.72	17
148	trichloroethene	11.62	11.81	11.44	11.96	12,14	213	*trans*,*trans*-2,4-hexadiene	9.87	9.75	9.96	9.89	17
149	tetrachloroethene	12.78	12.06	12.30	12.29	12	214	4-methyl-1,3-pentadiene	9.88	9.73	9.85	9.97	17
150	*cis*-1,3-dichloropropene	11.08	11.63	11.24	11.33	12,14	215	*trans*-1,3-hexadiene	9.95	9.89	9.95	9.88	17
151	*trans*-1,3-dichloropropene	10.84	11.60[p]	11.20[p]	11.34[p]	12,14	216	*trans*-1,3-pentadiene	10.00	10.03	10.05	9.99	17
152	3-chloro-1-propene	10.77	11.42	11.17	10.84	12,14	217	*trans*-1,4-hexadiene	10.04	10.32	10.18	9.87	17
153	2-(chloromethyl)-3-chloro-1-propene	10.47	11.55	10.68	10.69	12,14	218	Δ3-carene	10.06	9.99	10.10	10.22	17
							219	β-pinene	10.10	10.02	10.20	10.18	17
154	methyl vinyl ketone	10.73	11.36[p]	10.96[p]	10.56[p]	12	220	*trans*-4-octene	10.16	10.09	10.22	10.05	17
155	allyl aldehyde	10.71	11.33	10.90[cv]	10.87[cv]	12	221	*trans*-2-heptene	10.17	10.23	10.21	10.06	17
156	*trans*-2-butenal	10.44	11.18	10.67	10.70	12	222	*trans*-2-pentene	10.18	10.46	10.23	10.17	17
157	2-methyl-2-propenal	10.51	11.03	10.28	10.56	12	223	α-pinene	10.27	9.94	10.14	10.23	17
158	*cis*-3-hexene-2,5-dione	10.20	10.99	9.97	10.07	12	224	camphene	10.27	10.10[p]	10.26[p]	10.18[p]	17
159	*trans*-3-hexene-2,5-dione	10.28	11.26	10.30	10.27	12	225	1-heptene	10.39	10.50[p]	10.41[p]	10.40[p]	17
160	methyl vinyl ether	10.47	10.26	10.26	10.56	12,14	226	1-hexene	10.43	10.59	10.39	10.43	17
161	ethenone	10.77	10.65	10.33[cv]	10.56[cv]	12	227	1-pentene	10.50	10.73[p]	10.39[p]	10.46[p]	17
162	1-propen-1-one	10.16	10.21[p]	10.15[p]	10.02[p]	12	228	3,3-dimethyl-1-butene	10.55	10.98	10.43	10.55	17
163	1-buten-1-one	9.93	10.13	10.00	9.96	12	229	3-methyl-1,2-butadiene	10.24	9.31	10.20	10.38	17
164	2-methyl-1-propen-1-one	9.97	9.99	10.05	9.89	12	230	1,2-pentadiene	10.45	9.35	10.35	10.36	17
165	ethene	11.07	11.47	10.98	11.11	14	231	1,2-butadiene	10.58	9.42	10.56	10.51	17
166	propene	10.57	11.11	10.61	10.66	14	232	propadiene	11.01	9.51	11.03	11.10	17
167	*cis*-2-butene	10.25	10.56	10.31[cv]	10.35[cv]	14,17	233	benzene	11.85	11.54	12.25[cv]	11.63[cv]	14
168	*trans*-2-butene	10.19	10.52	10.29	10.33	14,17	234	toluene	11.22	11.09	11.28	11.20	14
169	1-butene	10.50	10.78	10.45	10.52	14,17	235	*o*-xylene	10.89	10.80	10.87	10.77	14
170	2-methylpropene	10.29	10.81	10.40	10.41	14,17	236	*m*-xylene	10.64	10.81	10.81	10.76	14
171	3-methyl-1-butene	10.50	10.75	10.39	10.40	14,17	237	*p*-xylene	10.89	10.72	10.75	10.67	14
172	2-methyl-2-butene	10.06	10.29	10.07[cv]	10.12[cv]	14,17	238	1,2,3-trimethylbenzene	10.49	10.69	10.64	10.48	14
173	2,3-dimethyl-2-butene	9.96	10.16	9.99	9.96	14,17	239	1,2,4-trimethylbenzene	10.49	10.60	10.48	10.35	14
174	*cis*-2-pentene	10.18	10.49	10.25	10.19	17	240	1,3,5-trimethylbenzene	10.24	10.74	10.47	10.49	14
175	2-methyl-1-butene	10.22	10.36	10.13	10.34	17	241	ethylbenzene	11.22	10.93	11.23	11.10	14
176	2,3-dimethyl-2-pentene	10.01	9.95	9.90	9.90	14	242	*n*-propylbenzene	11.35	10.85	11.21[cv]	11.08[cv]	14
177	*trans*-4,4-dimethyl-2-pentene	10.26	10.60[p]	10.14[p]	10.52[p]	14	243	isopropylbenzene	11.30	10.91	11.24	11.41	14
178	*cis*-1,3-pentadiene	10.00	10.05	10.05	9.98	14	244	phenol	10.58	10.93	10.66	10.53	14
179	1,5-hexadiene	10.21	10.62	10.36	10.03	14,17	245	methoxybenzene	10.78	10.90	10.85	10.76	14
180	2,5-dimethyl-2,4-hexadiene	9.68	9.17	9.70	9.63	14,17	246	*o*-cresol	10.38	10.63	10.36	10.41	14
181	1,3-butadiene	10.18	10.25	10.16	10.19	17	247	*m*-cresol	10.19	10.66	10.35	10.47	14
182	2-methyl-1,3-butadiene	10.00	9.96	9.96	10.22	14,17	248	*p*-cresol	10.33	10.56	10.27	10.33	14
183	1,3-cyclohexadiene	9.79	10.00	9.97	9.84	14,17	249	*o*-ethyltoluene	10.96	10.64[p]	10.87[p]	10.68[p]	14
184	1,3-cycloheptadiene	9.86	9.98	9.92	9.79	14	250	*m*-ethyltoluene	10.74	10.66	10.88	10.69	14
185	1,4-pentadiene	10.27	10.63	10.31	10.10	14,17	251	*p*-ethyltoluene	10.96	10.58	10.79	10.59	14
186	2-methyl-1,4-pentadiene	10.10	10.19	10.06	10.09	14	252	fluorobenzene	12.16	11.74[p]	12.21[p]	11.72[p]	14
187	2-methyl-1,5-hexadiene	10.02	10.15	10.04	10.03	14,17	253	chlorobenzene	12.11	11.75	12.21[cv]	11.97[cv]	14
188	2,5-dimethyl-1,5-hexadiene	9.92	9.81	9.93	9.98	14,17	254	bromobenzene	12.11	11.72	12.09	11.99	14
189	3,7-dimethyl-1,6-octadiene	9.74	9.59[p]	9.81[p]	9.83[p]	14	255	iodobenzene	11.96	11.73	12.08	12.03	14
190	*trans*-1,3,5-hexatriene	9.95	9.66	9.58	9.86	14,17	256	benzotrifluoride	12.34	12.35	12.41	12.45	14
191	cyclopentene	10.17	10.61	10.21	10.08	14,17	257	aniline	9.95	10.25	9.82	9.69	14
192	cyclohexene	10.17	10.52[p]	10.21[p]	10.05[p]	14,17	258	*N*,*N*-dimethylaniline	9.83	10.37[p]	10.20[p]	9.78[p]	12,14
193	cycloheptene	10.13	10.48	10.26	10.22	14,17	259	benzonitrile	12.48	11.58	11.61	12.42	12
194	1-methyl-cyclohexene	10.03	10.06[p]	9.92[p]	9.98[p]	14	260	nitrobenzene	12.85	12.33	12.75	12.86	12
195	bicyclo[2.2.1]-2-heptene	10.31	10.48	10.20[cv]	10.11[cv]	14	261	4-chlorobenzotrifluoride	12.62	12.53	12.55	12.41	14
196	bicyclo[2.2.2]-2-octene	10.39	10.46	10.27	10.20	14	262	o-dichlorobenzene	12.38	12.02	12.29	12.28	12
197	1,4-cyclohexadiene	10.00	10.34	10.13	9.96	14,17	263	m-dichlorobenzene	12.14	12.11	12.38	12.28	12
198	bicyclo[2.2.1]-2,5-heptadiene	9.92	10.44	10.06	9.96	14,17	264	p-dichlorobenzene	12.38	12.00	12.19	12.20	14
199	α-phellandrene	9.50	9.34	9.54[cv]	9.62[cv]	14,17	265	*p*-chloroaniline	10.08	10.60	10.26	10.12	12
200	α-terpinene	9.44	9.11	9.48	9.58	14,17	267	1,2,4-trichlorbenzene	12.28	12.30[p]	12.71[p]	12.20[p]	12,14
201	D-limonene	9.77	9.59[p]	9.78[p]	9.70[p]	17	268	*n*-propylpentafluorobenzene	11.52	11.72	11.48	11.63	14
202	γ-terpinene	9.75	9.54	9.71	9.69	17	269	hexafluorobenzene	12.70	13.21	12.66	12.79	12,14
203	1,3,5-cycloheptatriene	10.01	9.66	9.98	9.72	14,17	270	biphenyl	11.14	10.94	11.27	10.91	14

Table 2. (Continued)

no.	compound	$-\log(k_{OH})^a$	type 1[b]	type 2[b]	type 3[b]	ref	no.	compound	$-\log(k_{OH})^a$	type 1[b]	type 2[b]	type 3[b]	ref
271	benzyl chloride	11.54	11.43	11.48	11.69	12,14	292	*trans*-β-methylstyrene	10.23	10.41	10.33	10.22	12,14
272	benzaldehyde	10.89	11.41	11.28	10.84	12	293	β-dimethylstyrene	10.48	10.16	10.22	10.24	14
273	thiophenol	10.95	11.10	10.77	11.14	14	294	naphthalene	10.67	10.77	10.68	10.50	14
274	2,3-dimethylphenol	10.10	10.47	10.17[cv]	10.24[cv]	14	295	1-methylnaphthalene	10.28	10.51	10.49	10.51	14
275	2,4-dimethylphenol	10.15	10.40	10.06	10.14	14	296	2-methylnaphthalene	10.28	10.53	10.49	10.54	14
276	2,5-dimethylphenol	10.10	10.46	10.10	10.22	14	297	2,3-dimethylnaphthalene	10.11	10.41	10.51	10.62	14
277	2,6-dimethylphenol	10.18	10.47	10.11	10.22	14	298	1,4-dimethylnaphthalene	11.24	10.38	10.44	10.50	14
278	3,4-dimethylphenol	10.09	10.40	10.12[cv]	10.16[cv]	14	299	fluorene	10.92	10.34	10.74	11.04	14
279	3,5-dimethylphenol	9.95	10.50	10.11	10.23	14	300	phenanthrene	10.51	10.69	10.39	10.59	14
280	1,4-benzodioxane	10.80	10.84	10.88	10.66	14	301	anthracene	9.95	10.31	9.90	9.94	14
281	2,3-dihydrobenzofuran	10.52	10.56	10.56	10.45	14	302	acenaphthene	10.10	9.91	10.22[cv]	10.56[cv]	14
282	2,3-dichlorophenol	11.78	11.65	11.85	11.88	14	303	acenaphthylene	9.96	10.27	10.01	9.89	14
283	2,4-dichlorophenol	11.97	11.58	11.58	11.71	14	304	1,4-dichloronaphthalene	11.24	11.34	11.32	11.25	14
284	2,4-diaminotoluene	9.72	9.50	9.60[cv]	9.67[cv]	14	305	1,4-naphthoquinone	11.51	11.13[p]	11.63[p]	9.48[p]	14
285	*o*-chlorobiphenyl	11.55	11.24[p]	11.50[p]	11.19[p]	14	306	pyridine	12.31	12.02[p]	12.74[p]	12.47[p]	14
286	*m*-chlorobiphenyl	11.28	11.27	11.50	11.24	14	307	1,3,5-triazine	12.82	13.24	13.13	12.71	14
287	*p*-chlorobiphenyl	11.41	11.20	11.39[cv]	11.15[cv]	14	308	furan	10.39	10.54	10.30	10.70	14
288	*tert*-butylbenzene	11.34	11.14	11.26	11.44	14	309	pyrrole	9.96	10.34	10.38	10.17	14
289	tetralin	10.46	10.43	10.55	10.54	14	310	thiophene	11.02	10.94	10.61	11.08	14
290	styrene	10.26	10.58	10.42	10.41	14	311	thiazole	11.85	11.58	11.55	11.59	14
291	α-methylstyrene	10.28	10.34	10.25	10.15	12,14	312	2,3-benzofuran	10.43	10.61	10.40	10.44	14

[a] Units of cm^3 $molecule^{-1}$ s^{-1}. [b] Supercript abbreviations: cv = cross-validation set members; p = prediction set members.

chosen at random and designated the basis vector. The angle between this basis vector and each individual vector in the pool was computed. The descriptor most orthogonal to the single basis vector was added to the basis vector set and used to define a plane with the original basis vector. For each descriptor remaining in the pool, the angle between the descriptor and the plane was calculated. The descriptor most orthogonal to the basis set was then added to the basis set. Then, the descriptor most orthogonal to the space defined by the three basis vectors was selected for addition to the basis set. This process was repeated until all descriptors were "ranked" on the basis of orthogonality.

Clearly, the ordering of descriptors is going to be dependent on the descriptor initially selected as the single basis vector. To eliminate the starting point dependency, five randomly selected descriptors were used as starting points, and ordered lists based on orthogonalization were compiled for each. For data set 1, a pool of 31 descriptors was deemed acceptable for subjective feature selection (0.6×52 training set compounds $\cong 31$). Therefore, the pool of descriptors used for subjective feature selection was composed of the 31 descriptors ranked most highly on all five ordered lists. Such a process should work to eliminate any starting point dependency of the orthogonalization procedure. Interestingly, it was noted that although the exact sequence of descriptors varied, descriptors appeared at similar (not identical) ranks in all five lists.

Linear Model Formation and Validation. The reduced pool of descriptors was evaluated using evolutionary optimization procedures such as genetic algorithm[25] and simulated annealing.[26−28] Descriptor subsets were examined to see if they could successfully map k_{OH} on the basis of linear regression. Models formed in this manner are termed type 1 models. For each model examined, rms error was computed for the training set compounds by comparing experimental k_{OH} values to calculated k_{OH} values. Descriptor *T*-values were monitored to ensure model coefficients were meaningful. Additionally, variance of inflation factors (VIFs) were calculated for each descriptor by regressing that descriptor against all others in the model. VIFs were calculated as $[1 - R^2]^{-1}$, where R is the multiple correlation coefficient and models were determined to be free of multi-collinearities if VIFs for all descriptors in the model were less than 10.

Models identified using the above criteria were further examined. Plots of calculated versus observed log k_{OH} values were generated and correlation coefficients calculated for each model formed to further examine the relationship between calculated and observed k_{OH} values. Potential outliers were identified by examining residuals, studentized residuals, standardized residuals, leverage, DFFITS statistic, and Cook's distance.[42] Any observation failing more than three of the six tests was flagged as a potential outlier. Model parameters were investigated with potential outliers present and absent to determine their influence. On the basis of these comparisons, a decision was made as to whether or not the observation in question should be eliminated.

Models with varying numbers of descriptors were investigated. Initial models were found with relatively few descriptors, usually three. The number of descriptors was increased by one, and new models formed. Optimal model size was determined on the basis of rms error and correlation coefficient. Once a model was chosen as optimal, prediction set rms error was examined to demonstrate the ability of the model to accurately predict k_{OH} values for compounds not used in training. The descriptor subset defining the most predictive, statistically valid type 1 model was saved for consideration in generating nonlinear models.

Nonlinear Model Formation and Validation. The descriptor subset forming the optimal type 1 model was used to generate a nonlinear CNN model. Such nonlinear models, built using descriptor subsets chosen by an evolutionary optimization procedure with a linear fitness evaluator, are termed type 2 models. The CNN was a three-layer (input, hidden, output), fully connected, feed-forward network. The number of neurons in the input layer was determined by the size of the descriptor subset defining the optimal type 1 model, and the output layer consisted of one neuron representing k_{OH}. The number of hidden layer neurons was

Table 3. Five-Descriptor Type 1 Linear Regression Model Selected by Evolutionary Optimization Techniques for Data Set 1

descriptor	coefficient	error estimate	range	explanation[a]
2SP2	-1.36×10^{-1}	1.5×10^{-2}	0−6	no. of secondary sp^2 carbons
3SP2	-4.15×10^{-1}	3.1×10^{-2}	0−3	no. of tertiary sp^2 carbons
MDE13	-6.70×10^{-2}	1.57×10^{-2}	0.00−7.08	distance edge for 1° and 3° carbons
MDE23	-4.07×10^{-2}	9.7×10^{-3}	0.00−8.10	distance edge for 2° and 3° carbons
MDE34	1.08×10^{-1}	2.5×10^{-2}	0.00−6.44	distance edge for 3° and 4° carbons
constant	1.07×10^{1}	1×10^{-1}		intercept

[a] 2SP2, number of sp^2 hybridized carbon atoms bonded to two other carbon atoms; 3SP2, number of sp^2 hybridized carbon atoms bonded to three other carbon atoms; MDE13, molecular distance edge term between primary and ternary carbons;[43] MDE23, molecular distance edge term between secondary and ternary carbons;[43] MDE34, molecular distance edge term between ternary and quaternary carbons.[43]

varied, and rms error was examined for the training and cross-validation sets. The point at which an additional hidden layer neuron did not reduce rms error of the training and cross-validation sets defined the most appropriate architecture. Fully trained networks were used in conjunction with the external prediction set to demonstrate the ability of the network to generalize.

The final model type examined, type 3 models, are the most computationally intensive. The reduced pool of descriptors was submitted to a genetic algorithm, and a CNN was used as the fitness evaluator. Descriptor subset fitness was again based on training and cross-validation set rms error. Selected subsets were used to fully train a three-layer network, and architecture was determined as described above. After network training was complete, prediction set compounds were used to verify the ability of the network to accurately predict k_{OH} values for compounds not involved in training.

RESULTS AND DISCUSSION

Data Set 1. Evolutionary optimization procedures were used to screen many descriptor subsets. The smallest subset of descriptors with the lowest rms error and highest correlation coefficient was selected as the final type 1 model. A five-descriptor model with a training set rms error of 0.106 log units and $R = 0.932$ was selected as optimal. All *T*-values were greater than 4, and no VIF exceeded 10. Calculated k_{OH} values for all 57 compounds are shown in Table 1. Two compounds, 1,3,5-cycloheptatriene (23) and propadiene (57), were identified as potential outliers on the basis of their descriptor values. Model coefficients were calculated with and without the identified compounds to determine their influence. Removal of the compounds caused model coefficients to change by an average of 20%. However, the prediction errors for the two compounds were within the range of errors for the other 50 training set compounds, and removal of the two potential outliers increased the rms error to 0.111 log units and decreased *R* to 0.926. Additionally, the data set was already quite small, making it difficult to justify removing compounds, and all 57 compounds were used in other studies.[17] Therefore, the two potential outliers were left in the data set. Using the five-descriptor model based on all 52 training set compounds, the rms error for the five prediction set compounds was 0.139 log units.

The five descriptors used in the best type 1 model are shown in Table 3. All five of the descriptors are topological, which is advantageous since such descriptors do not require geometry optimization of compounds. Among the five descriptors, pairwise correlation ranged from 0.008 to 0.769 with an average value of 0.316. Two of the descriptors, 2SP2 and 3SP2, represent counts of particular sp^2 hybridized carbon types. 2SP2 denotes sp^2 hybridized carbons bonded to two other carbon atoms, and 3SP2 counts sp^2 hybridized carbons bonded to three other carbon atoms. Both descriptors probably encode information concerning attack sites for the radical.

The other three descriptors, MDE13, MDE23, and MDE34, are molecular distance edge descriptors.[43] These descriptors are also concerned with carbon atoms. Carbon atoms are divided into four types: (1) primary ($-CH_3$), (2) secondary ($>CH_2$), (3) ternary ($>CH-$), and (4) quaternary ($>C<$). A distance edge term is computed for all pairwise combinations of carbon types, for a total of 10 descriptors. For example, MDE13 represents the distance edge descriptor between primary and ternary carbons. In addition to describing carbon bonding, these descriptors include information regarding distance between atoms.[43] The three distance edge descriptors probably encode branching information and steric considerations.

Although it is surprising that the best model resulted from five topological descriptors, it is not surprising that all five descriptors are derived from carbon atom information. The reaction center for radical reactions is often an unsaturated carbon. And, since data set 1 contains only unsaturated hydrocarbons, such descriptors can adequately describe the data set. As the diversity of the data set increases, e.g., with the addition of compounds containing heteroatoms, reactions will be less simplistic and descriptors with different information will be necessary.

As stated earlier, rms error was 0.106 log units for the training set and 0.139 log units for the prediction set. The residual sum of squares (RSS) was also computed for all 57 compounds in data set 1. This was done to facilitate comparison with the work in ref 17. Table 1 shows that the RSS of 0.676 log units for the type 1 model compares favorably with the PLS model of Medven (RSS = 0.710 log units) and Klamt's QSPR (RSS = 0.781 log units). Atkinson's fragment contribution method has a lower RSS (0.382 log units) than the type 1 model.

The five descriptors forming the type 1 linear model were used to build a nonlinear type 2 CNN model. The input layer of the network consisted of five neurons, one for each of the descriptors in Table 3. The output layer contained a single neuron representing predicted k_{OH} values. The number of hidden layer neurons was varied to find the optimal architecture. To begin with, two hidden neurons were used, and the network was trained. An additional hidden layer

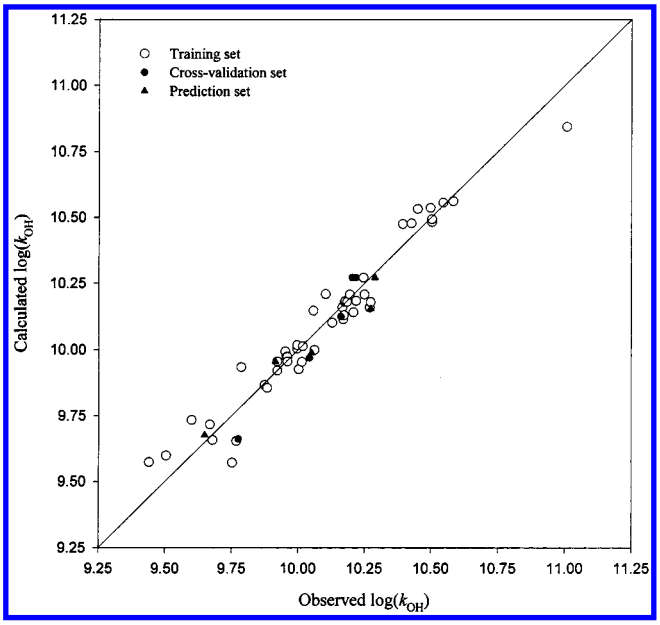

Figure 1. Plot of calculated vs observed log(k_{OH}) of the training, cross-validation, and prediction set compounds using the type 2 model for data set 1. Table 1 shows the calculated and observed values for each compound.

neuron was added, the new network was trained, and networks were compared using rms errors for the training and cross-validation sets. The process was repeated until additional hidden layer neurons did not enhance network performance. The optimal network architecture is that which produces the lowest rms errors with the fewest adjustable parameters. The number of adjustable parameters is computed as AP = IL × HL + HL × OL + HL + OL, where AP represents the number of adjustable parameters and IL, HL, and OL denote the number of neurons in the input layer, hidden layer, and output layer, respectively. It should be noted that the ratio of training set observations to adjustable parameters should be kept above 2.0 to avoid overtraining.

Network training involved adjusting weights and biases to minimize rms error. The quasi-Newton method BFGS (Broyden−Fletcher−Goldfarb−Shanno)[44] was used to train the network. During training, the rms error of the cross-validation set was computed periodically. The point at which the cross-validation rms error began to increase defined the stopping point for training. Further training of the CNN would result in memorization of idiosyncrasies of the training data, reducing the ability of the network to generalize to compounds not used in training.

For data set 1, a 5−2−1 neural network architecture was selected as optimal. This model resulted in rms errors of 0.0705 log units and 0.0741 log units for the training and cross-validation sets, respectively. The CNN was able to generalize well, as shown by the rms error of 0.0639 log units for the external prediction set. Figure 1 shows a plot of calculated versus observed k_{OH} values for the training, cross-validation, and prediction sets for results generated with the type 2 model. The one-to-one correlation line clearly demonstrates that the model is accurately predicting k_{OH} values.

Table 1 contains the estimated value for each of the 57 compounds. Compounds 23 and 57, flagged as potential outliers during linear model formation, are more accurately predicted with the type 2 model compared to the type 1 model. Additionally, the RSS error using the type 2 model was reduced to 0.281 log units, a 58% improvement relative to the type 1 model. This also represents a 26% improvement relative to the fragment contribution method of Atkinson. Recall that the five descriptors used in this model are topological, so no geometry optimization is required. Predictions of future compounds using this model require only simple, two-dimensional sketches of the compound of interest.

Table 4. Five Descriptors Forming the Type 3 Model for Data Set 1

descriptor	range	explanation[a]
PND	0 to 29.31	superpendentic index
MCB	1 to 3	multiple carbon−carbon bonds
LUMO	−1.097 to 1.389	lowest unoccupied molecular orbital
ENEG	3.87 to 4.42	electronegativity
FPSA3	0.06738 to 0.1773	fractional positive surface area

[a] PND, square root of the sum of the products of all nonzero elements of the pendent matrix;[45] MCB, number of multiple carbon−carbon bonds (aromatic rings do contribute to this count); LUMO, energy of the lowest unoccupied molecular orbital obtained from MOPAC output when an AM1 Hamiltonian was used for geometry optimization; ENEG, average of HOMO and LUMO energies obtained from MOPAC output when an AM1 Hamiltonian was used for geometry optimization; FPSA3, fractional positively charged surface area = (sum of (positively charged partial surface area × partial positive charge)) ÷ (total molecular surface area).[40]

A type 3 model was also built for data set 1. The 31 descriptors in the reduced pool were submitted to a genetic algorithm with a 5−2−1 CNN as the fitness evaluator. The fitness of descriptor subsets was calculated as COST = TSET + 0.4 × |TSET − CVSET|, where TSET and CVSET denote rms errors for the training and cross-validation sets, respectively. This fitness function gives preference to descriptor subsets that produce low, equivalent rms errors for the training and cross-validation sets. Past work in this laboratory has demonstrated that networks producing low COST values will generalize well. That is, CNNs that produce training and cross-validation set rms errors that are low and similar in magnitude tend to perform well in predicting properties of interest for compounds not used in the training process.

The five descriptors forming the best type 3 model are shown in Table 4. Pairwise correlations for the five descriptors ranged from 0.104 to 0.736, with an average value of 0.292. Of the five descriptors selected, two are topological (PND and MCB), two are electronic (LUMO and ENEG), and one is a hybrid descriptor (FPSA3). PND denotes the superpendentic index,[45] which measures the degree of branching. MCB is simply a count of multiple carbon−carbon bonds. The electronic descriptor LUMO represents the energy of the lowest unoccupied molecular orbital of the compound, and ENEG is the average energy of the highest occupied molecular orbital and lowest unoccupied molecular orbital. These two descriptors encode the energetics of the reactant molecular orbitals that will be involved in the reaction. The hybrid FPSA3 descriptor represents a charge weighted fractional positive surface area. Electronic information regarding energetics of reactants is encoded in this descriptor.

With the type 3 model, training set rms error is reduced to 0.0638 log units and cross-validation set rms error is

Table 5. Ten-Descriptor Type 1 Linear Regression Model Selected by Evolutionary Optimization Techniques for Data Set 2

descriptor	coefficient	error estimate	range	explanation[a]
KAPA3	-5.66×10^{-2}	1.34×10^{-2}	0.000 to 9.143	${}^3\kappa$ index
MOLC5	-1.82×10^{-1}	3.9×10^{-2}	0 to 4	path three molecular connectivity
NAB	1.32×10^{-1}	1.2×10^{-2}	0 to 16	no. aromatic bonds
NLP	-2.27×10^{-1}	1.9×10^{-2}	0 to 8	no. lone pairs
WTPT3	1.37×10^{-1}	1.6×10^{-2}	0.00 to 14.36	sum of weighted paths starting from heteroatoms
MDE14	1.13×10^{-1}	2.3×10^{-2}	0 to 4	distance edge for 1° and 4° carbons
3SP2	-2.56×10^{-1}	4.9×10^{-2}	0.000 to 8.485	no. of tertiary sp^2 carbons
PND	-1.14×10^{-2}	2.9×10^{-3}	0 to 132	superpendentic index
LUMO	-7.91×10^{-1}	5.0×10^{-2}	−1.919 to 4.159	lowest unoccupied molecular orbital
ENEG	1.45	0.06	3.205 to 8.002	electronegativity measure
constant	3.99	0.29		intercept

[a] KAPA3, ${}^3\kappa = 4$(max no. 3 − bond fragments)(min no. 3 − bond fragments)/(actual no. 3 − bond fragments), κ index based on path lengths of three;[33] MOLC5, path three molecular connectivity;[34] NAB, number of aromatic bonds; NLP, number of lone pairs; WTPT3, sum of all path weights starting from heteroatoms;[46] MDE14, molecular distance edge term between primary and quaternary carbons;[43] 3SP2, number of sp^2 hybridized carbon atoms bonded to three other carbon atoms; PND, square root of the sum of the products of all nonzero elements of the pendent matrix;[45] LUMO, energy of the lowest unoccupied molecular orbital obtained from MOPAC output when an AM1 Hamiltonian was used for geometry optimization; ENEG, average of HOMO and LUMO energies obtained from MOPAC output when an AM1 Hamiltonian was used for geometry optimization.

reduced to 0.0648 log units. The model generalizes well, as demonstrated by the prediction set rms of 0.0657 log units. Relative to the type 2 model, training and cross-validation rms errors improve, while prediction set rms error is slightly worse. Table 1 shows that the RSS error using the type 3 model is 0.234 log units. Although the type 3 model offers some overall improvement, the type 2 model is more attractive computationally. Descriptors used in the type 2 model do not require geometry optimized structures and are therefore much less computationally intense.

Data Set 2. Table 2 shows the 312 compounds, and corresponding log k_{OH} values, for the compounds forming data set 2. A significant amount of structural diversity is present, including functional groups such as alcohols, amines, sulfides, aromatics, and halogenated compounds. Rate constants cover a range of over 6 orders of magnitude. All compounds in data set 1 are also present in data set 2.

For data set 2, screening descriptor subsets using evolutionary optimization procedures resulted in selection of a ten-descriptor type 1 model. The rms error of the training data was 0.392 log units, and R was 0.936. All T-values were greater than 4, and no VIF factor exceeded 10. Calculated log k_{OH} values are shown in Table 2. One compound, *o*-nitrophenol (266), failed four of the six outlier tests, identifying it as a potential outlier. Model coefficients were examined with and without *o*-nitrophenol present during model formation. Removing *o*-nitrophenol changed the coefficients by an average of 1.4%, with all coefficients changing by less than 3%. Therefore, the compound was not removed from the data set. With this model, the rms error for the external prediction set was 0.319 log units, clearly demonstrating the ability of the model to generalize to compounds not used in training.

Table 5 shows the 10 descriptors selected for the type 1 model. Pairwise correlations for the 10 descriptors ranged from 0.008 to 0.841 with an average value of 0.249. Eight of the descriptors are topological, and two are electronic. KAPA3 denotes the path 3 κ index, which compares a molecule containing N atoms with the most and least branched forms of an N atom molecule. This descriptor encodes molecular shape. MOLC5 describes path three molecular connectivity for a molecule. The simple descriptors NAB and NLP are counts of aromatic bonds and lone pairs, respectively. WTPT3 denotes the sum of all weighted paths starting from heteroatoms. The molecular distance edge descriptor, MDE14, is different than the three selected for the data set 1 type 1 model in that it is concerned with primary and quaternary carbons. MDE14 probably encodes information about branching and steric factors. The four remaining descriptors were selected in previous models. The topological descriptor 3SP2 counts sp^2 hybridized carbons bonded to three other carbon atoms. This descriptor probably provides information regarding attach sites for radicals. PND represents the superpendentic index and encodes information on branching of the molecules. The electronic descriptors LUMO and ENEG contain information regarding HOMO and LUMO energies and provide insight into the energetics of the reactant molecule.

The rms errors for data set 2 are roughly 3.5 times larger than those for data set 1 when comparing type 1 models. Clearly, data set 2 is much more structurally diverse, making the model more complex. That is also why the number of descriptors necessary to form the data set 2 model was larger than the number necessary for the data set 1 model. But, the data set 2 model offers the distinct advantage of being applicable to any compound falling into one of the four described reaction classes, where all data set 1 models are strictly limited to unsaturated hydrocarbons. It is not at all surprising that a model focused on a narrow, similar subset of compounds produces rms errors lower than those produced by models encompassing a larger range of compounds with greater structural diversity. Typically, the application of interest will determine which type of model is more appropriate.

The 10 descriptors forming the type 1 model were used to generate type 2 models with various architectures. A 10−7−1 network architecture was selected as optimal. This network contains 85 adjustable parameters, corresponding to a ratio of 2.9 for training set observations (250) to adjustable parameters. This is well above the minimum acceptable ratio of 2.0.

With this model, the training set rms error was reduced to 0.229 log units, and cross-validation set rms error was 0.237 log units. The prediction set rms error was 0.254 log

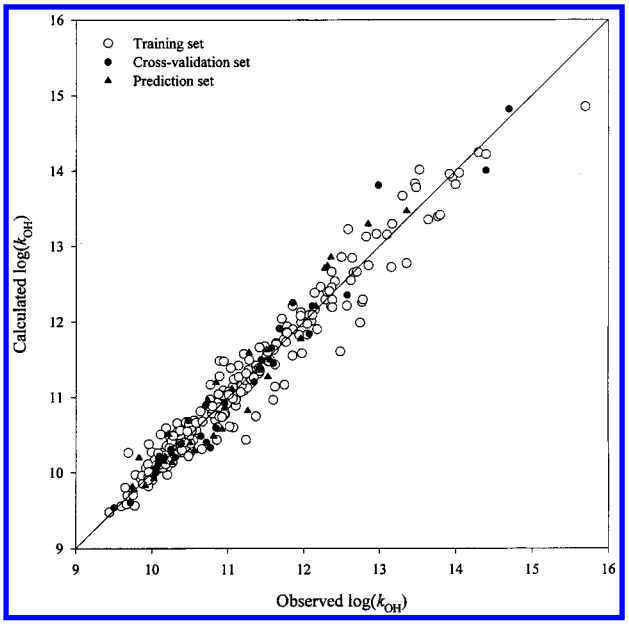

Figure 2. Plot of calculated vs observed log(k_{OH}) of the training, cross-validation, and prediction set compounds using the type 2 model for data set 2. Table 2 shows the calculated and observed values for each compound.

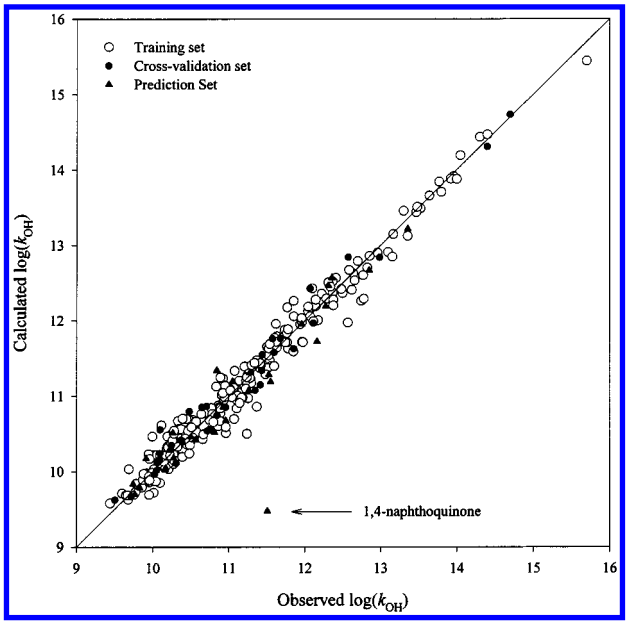

Figure 3. Plot of calculated vs observed log(k_{OH}) for the training, cross-validation, and prediction set compounds using the type 3 model for data set 2. Table 2 shows the calculated and observed values for each compound. Note that 1,4-naphthoquinone is the only quinone in the data set, and removal of this compound reduces the prediction set rms error over 50% to 0.202 log units.

units, showing the ability of the CNN to generalize to compounds not used in model formation. Compared to the type 1 model, rms errors are reduced by nearly half. The type 2 model is clearly a superior model.

Figure 2 shows a calculated versus observed log k_{OH} plot for the training, cross-validation, and prediction sets using the type 2 model. This plot clearly demonstrates that the CNN is describing a true relationship between molecular structure and log k_{OH}. Table 2 contains the predicted k_{OH} values for each of the 312 compounds. Together, Figure 2 and Table 2 illustrate the diverse range of compounds that can be accurately described by a single model.

The type 2 model formed for data set 2 was used to predict the log k_{OH} values for the 57 compounds in data set 1. Recall that all 57 compounds were present in data set 2, with some in each of the training, cross-validation, and prediction sets. For all 57 compounds, the rms error comes out to 0.0825 log units. This corresponds to an RSS of 0.388 log units. As seen in Table 1, this model produces an RSS lower than that produced by the PLS method or the QSPR of Klamt, while the RSS for Atkinson's fragment contribution method is slightly lower. The error calculated using the data set 2 model to predict the data set 1 compounds is higher than the error using data set 1 models for data set 1 compounds. This is because data set 2 is much more diverse and data set 1 is tailored to a specific type of compound. This finding suggests that, if accuracy is the primary focus, a model could be built using a smaller set of structurally similar compounds. But, if future unknowns will encompass a large range of functionalities, a model based on a diverse set of compounds will be more appropriate and will provide acceptable prediction accuracy.

The reduced pool of 109 descriptors was submitted to a genetic algorithm with a CNN fitness evaluator to form a type 3 model. A 10−8−1 network was selected, and the best model had a training set rms error of 0.173 log units and a cross-validation set rms error of 0.187 log units. A plot of calculated versus observed log k_{OH} values is shown in Figure 3. The error values are slightly lower than those obtained with the type 2 model. The 10 descriptors selected are shown in Table 6. Six of the descriptors are topological, one geometric, two electronic, and one hybrid. Pairwise correlations ranged from 0.006 to 0.727 with an average value of 0.252.

Table 6. Ten Descriptors Forming the Type 3 Model for Data Set 2

descriptor	range	explanation[a]
NC	0 to 14	no. of carbon atoms
NSB	0 to 12	no. of single bonds
NDB	0 to 3	no. of double bonds
WTPT3	0.00 to 14.36	sum weighted paths starting from heteroatoms
1SP2	0 to 2	no. of primary sp^2 carbons
MDE33	0.00 to 41.03	distance edge for 3° carbons
GEOM3	-1.11×10^{-16} to 0.9467	third major dimension
HOMO	−13.3 to −7.543	highest occupied molecular orbital
HARD	2.998 to 6.386	hardness measure
CHAA3	−0.00367 to 0.004778	sum of charge on acceptor atoms

[a] NC, number of carbon atoms; NSB, number of single bonds (aromatic rings do not contribute to this count); NDB, number of double bonds (aromatic rings do not contribute to this count); WTPT3, sum of all path weights starting from heteroatoms;[46] 1SP2, number of sp^2 hybridized carbon atoms bonded to one other carbon atom; MDE33, molecular distance edge term between pairs of ternary carbon atoms;[43] GEOM3, third principal dimension (thickness) of a molecule; HOMO, energy of the highest occupied molecular orbital obtained from MOPAC output when an AM1 Hamiltonian was used for geometry optimization; HARD, half the difference of the HOMO and LUMO energies obtained from MOPAC output when an AM1 Hamiltonian was used for geometry optimization; CHAA3, (sum of charges on hydrogen-bond acceptor atoms) ÷ (total molecular surface area).

NC, NSB, and NDB are counts of carbon atoms, single bonds, and double bonds, respectively. WTPT3, which was also present in the type 1 model for data set 2, denotes the sum of all path weights starting from heteroatoms. The molecular distance edge descriptor describing ternary carbons

(MDE33) probably encodes information on degree of branching. 1SP2 counts the number of sp^2 hybridized carbon atoms bonded to only one other carbon and probably provides information on attack sites for radicals. GEOM3 is a geometric descriptor that provides the third major moment, or thickness, of a molecule. The electronic descriptors HOMO and HARD contain information about molecular orbitals. HOMO denotes the energy of the highest occupied molecular orbital, and HARD is half the difference of the HOMO and LUMO energies; i.e., HARD = 0.5 × (HOMO − LUMO). The hybrid descriptor CHAA3 encodes information about hydrogen bonding and is calculated as the sum of charges on acceptor atoms divided by the molecular surface area.

The type 3 model described above produces a prediction set rms error of 0.416 log units. This value is significantly larger than the training and cross-validation set errors. However, examination of Figure 3 reveals that this error is largely the result of one compound. The calculated value for 1,4-naphthoquinone (305) is much lower than the actual value resulting in a large deviation from the line of one-to-one correspondence. This compound is the only quinone in the data set which may explain the poor prediction. However, the type 1 and 2 models did not have this problem. Removing 1,4-naphthoquinone lowers the prediction set rms error to 0.202 log units which is in line with the training and cross-validation sets.

To verify that models formed were not simply based on chance correlations, a Monte Carlo experiment was conducted. The k_{OH} values for the 312 compounds in data set 2 were randomized. The reduced pool of 109 descriptors was fed to a genetic algorithm with a 10−8−1 CNN as the fitness evaluator to form a type 3 model. The resulting model produced a training set rms error of 1.051 log units, a cross-validation set rms error of 1.120 log units, and a prediction set rms error of 1.069 log units. This demonstrates that the models developed earlier in this paper accurately describe a true relationship between molecular structure and k_{OH}.

CONCLUSIONS

For data set 1, a five-descriptor type 2 model was built using linear methods of feature selection and a 5−2−1 CNN for model building. All descriptors used were topological and required only a two-dimensional sketch of the molecules. Resulting rms errors were 0.0705 log units, 0.0741 log units, and 0.0639 log units for the training, cross-validation, and prediction sets, respectively. The RSS for all 57 compounds was 0.281 log units, which is a 26% improvement compared to the 0.382 log units achieved with Atkinson's fragment contribution method. On the basis of this small set of data, the QSPR described in this paper appears to be the most effective in predicting hydroxyl radical rate constants.

Ten-descriptor models were generated on the basis of the diverse set of 312 compounds comprising data set 2. This data set allowed examination of the potential of developing a single model to predict all unknowns as opposed to developing several models, one for each class of compounds encountered. A 10−7−1 type 2 CNN was developed that produced a training set rms = 0.229 log units, a cross-validation set rms = 0.237 log units, and a prediction set rms = 0.254 log units. Using this model, the RSS for the data set 1 compounds was 0.388 log units, which still compares well with existing methodologies.

REFERENCES AND NOTES

(1) Atkinson, R. Gas-Phase Reactions of the Hydroxyl Radical. *Chem. Rev.* **1986**, *86*, 69−201.
(2) Atkinson, R. Kinetics and Mechanisms of the Gas-Phase Reactions of the Hydroxyl Radical with Organic Compounds. *J. Phys. Chem. Ref. Data Monogr. 1* **1989**, 1−216.
(3) Atkinson, R. Gas-Phase Tropospheric Chemistry of Organic Compounds. *J. Phys. Chem. Ref. Data Monogr. 2* **1994**, 1−246.
(4) Heicklen, J. The Correlation of Rate Coefficients for H-Atom Abstraction by HO Radicals with C-H Bond Dissociation Enthalpies. *Int. J. Chem. Kinet.* **1981**, *13*, 651−665.
(5) Cohen, N. The Use of Transition-State Theory to Extrapolate Rate Coefficients for Reactions of OH with Alkanes. *Int. J. Chem. Kinet.* **1982**, *14*, 1339−1362.
(6) Jolly, G. S.; Paraskevopoulos, G.; Singleton, D. L. Rate of OH Radical Reactions. XII. The Reactions of OH with *c*-C_3H_6, *c*-C_5H_{10}, and *c*-C_7H_{14}. Correlation of Hydroxyl Rate Constants with Bond Dissociation Energies. *Int. J. Chem. Kinet.* **1984**, *17*, 1−10.
(7) Cohen, N.; Benson, S. W. Empirical Correlations for Rate Coefficients for Reactions of OH with Haloalkanes. *J. Phys. Chem.* **1987**, *91*, 171−175.
(8) Gaffney, J. S.; Levine, S. Z. Predicting Gas Phase Organic Molecule Reaction Rates Using Linear Free-Energy Correlations. I. O(^{3}P) and OH Addition and Abstraction Reactions. *Int. J. Chem. Kinet.* **1979**, *11*, 1197−1209.
(9) Rinke, M.; Wahner, A.; Zetzsch, C. Dependence of the Rate of OH Addition to Aromatics on the Ionization Potential: A Predictive Tool for Rate Constants. *J. Photochem.* **1981**, *17*, 142.
(10) Darnall, K. R.; Atkinson, R.; Pitts, J. N., Jr. Rate Constants for the Reaction of the OH Radical with Selected Alkanes at 300 K. *J. Phys. Chem.* **1978**, *82*, 1581−1584.
(11) Atkinson, R. Estimations of OH Radical Rate Constants from H−Atom Abstraction from C−H and O−H Bonds over the Temperature Range 250−1000 K. *Int. J. Chem. Kinet.* **1986**, *18*, 555−568.
(12) Atkinson, R. A Structure−Activity Relationship for the Estimation of Rate Constants of OH Radicals with Organic Compounds. *Int. J. Chem. Kinet.* **1987**, *19*, 799−828.
(13) Atkinson, R. Estimation of Gas-Phase Hydroxyl Radical Rate Constants for Organic Chemicals. *Environ. Toxicol. Chem.* **1988**, *7*, 435−442.
(14) Klamt, A. Estimation of Gas-Phase Hydroxyl Radical Rate Constants of Organic Compounds from Molecular Orbital Calculations. *Chemosphere* **1993**, *26*, 1273−1289.
(15) Klamt, A. Estimation of Gas-Phase Hydroxyl Radical Rate Constants of Oxygenated Compounds Based on Molecular Orbital Calculations. *Chemosphere* **1996**, *32*, 717−726.
(16) Güsten, H.; Medven, Z.; Sekušak, S.; Sabljic, A. Predicting Tropospheric Degradation of Chemicals: From Estimation to Computation. *SAR QSAR Environ. Res.* **1995**, *4*, 197−209.
(17) Medven, Z.; Güsten, H.; Sabljic, A. Comparative QSAR Study on Hydroxyl Radical Reactivity with Unsaturated Hydrocarbons: PLS Versus MLR. *J. Chemom.* **1996**, *10*, 135−147.
(18) Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.
(19) Wessel, M. D.; Jurs, P. C.; tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(20) Engelhardt McClelland, H.; Jurs, P. C. Prediction of Vapor Pressure of Organic Compounds Using a Quantitative Structure−Property Relationship. Manuscript in preparation.
(21) Johnson, S. R.; Jurs, P. C. Prediction of Liquid Crystal Clearing Temperatures from Molecular Structure. *Chem. Mater.* **1999**, *11*, 1007−1023.
(22) Bakken, G. A.; Jurs, P. C. Prediction of Methyl Radical Addition Rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 508−514.
(23) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(24) Jurs, P. C.; Chow, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; pp 103−129.
(25) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(26) Kalivas, J. H.; Roberts, N.; Sutter, J. M. Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet−Visible Spectrophotometry. *Anal. Chem.* **1989**, *61*, 2024−2030.
(27) Kalivas, J. H. Generalized Simulated Annealing for Calibration Sample Selection from an Existing Set and Orthogonalization of Undesigned Experiments. *J. Chemom.* **1991**, *5*, 37−48.
(28) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(29) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure-Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(30) Stewart, J. P. P. *Mopac 6.0*, *Quantum Chemistry Program Exchange*; Indiana University, Bloomington, IN; Program 455.
(31) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(32) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5−13.
(33) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.
(34) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press, Ltd.: Hertfordshire, England, 1986.
(35) Bondi, A. van der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441−451.
(36) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.
(37) Miller, K. J.; Savchik, J. A. A New Empirical Method To Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.
(38) Abraham, R. J.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Comput. Chem.* **1987**, *9*, 288−297.
(39) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.
(40) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(41) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(42) Belsley, D. A.; Kuh, E.; Welson, R. E. *Regression Diagnostics*; Wiley: New York, 1980.
(43) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(44) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.
(45) Madan, A. K.; Gupta, S.; Singh, M. Superpendentic Index: A Novel Highly Discriminating Topological Descriptor for Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 272−277.
(46) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *4*, 162−175.

CI990042A