Prediction of Methyl Radical Addition 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 March 24, 1998

Multiple linear regression and computational neural networks (CNNs) are used to develop quantitative structure−property relationships for methyl radical addition rate constants. Structure based descriptors are used to numerically encode substrate information for 191 compounds. Descriptors can be classified as topological, geometric, electronic, or combination. A six-descriptor CNN was developed that produced training set rms error = 0.381 log units and rms error = 0.496 log units for an external prediction set. A seven-descriptor CNN was used to build a model for a subset of 172 of the compounds. Training set rms error was 0.424 log units and prediction set rms error reduced to 0.409 log units. Model predictions were on the order of experimental error.

INTRODUCTION

Methyl radicals are important in many areas of chemistry including polymerizations[1] and atmospheric reactions.[2] The main source of methyl radicals in the atmosphere is from the reaction of methane with hydroxyl radicals. Methane is the second most important greenhouse gas after carbon dioxide, and its fate is relevant to global warming and ozone production and loss.[2] Since methyl radical reactions control the fate of most atmospheric methane, the rates of these reactions are extremely important parameters.

Experimental determination of rate constants for methyl radical addition dates back to early work by Szwarc,[3] who measured methyl radical affinities. Recent work has focused on confirming and expanding Szwarc's results[4,5] as well as converting the relative methyl radical affinities to absolute rate constants. Current methodology involves using ESR spectroscopy to monitor how the radical of interest interacts with particular substrates.[4,5] Details of experimental conditions for all data presented in this study can be found in ref 5 and references therein. Significant amounts of time and money are required to determine rate constants experimentally.

In addition to experimental studies, much theoretical work has been devoted to the methyl radical.[6−11] Ab initio studies have been done to examine activation energies, transition states, and heats of reaction.[7−9,11] Results have been used to obtain rate constants for certain reactions, which agree fairly well with experiment. Additionally, methods to estimate reactivity ratios, most notably Alfrey-Price Q-e parameters,[12−14] have been developed.[15,16] But, such methods require some experimental work for each substrate.

This paper presents results obtained for quantitative structure−property relationships (QSPRs) for methyl radical addition rate constants (k_r). Related models have been successfully developed for prediction of hydroxyl radical addition rate constants.[17−20] Multiple linear regression and computational neural networks (CNNs) are used to relate substrate structure to k_r. Numerical descriptors are calculated in order to encode structural features. Descriptors are selected to build models to accurately predict addition rate constants.

EXPERIMENTAL SECTION

The k_r values for the small organic molecules used in this study were compiled from literature data.[3,21−47] Invaluable assistance in collecting the data was provided by Dr. Kàroly Héberger. All reactions are believed to be by addition only.[5] Data selected were for reactions run at 338 K. Clearly, varying temperature causes variations in rates of reaction. The compounds used are presented in Table 1, along with their experimentally determined k_r values. The compounds are grouped according to common structural features.

Of the 191 compounds used in this study (data set 1), 172 compounds formed the training set for linear models, and 19 were randomly selected and placed in an external prediction set. Good models were identified by minimization of root-mean-square (rms) error for the training set, with concurrent ability to generalize to the prediction set. That is, rms error of the training set was used to direct the search for a descriptor subset for a linear model, and compounds held out of training were used after model formation to validate the model and ensure predictive ability.

CNNs were used to build nonlinear models. First, a 19-member cross-validation set was removed from the training data, leaving 153 training set members, 19 cross-validation set members, and 19 prediction set members. The cross-validation set is required to prevent network over-training, and its use will be discussed more fully in the results and discussion section. Additional models were built using a subset of 172 compounds (data set 2) composed of a 155-member training set and 17 prediction set members. Data Set 2 was formed after removing compounds defined in Table 1 as allenes, alkynes, and heterocycles. The rationale for removing the compounds will be presented in the Results and Discussion section.

QSPR models were developed using the Automated Data Analysis and Pattern recognition Toolkit (ADAPT)[48,49] software system as well as genetic algorithm,[50] simulated annealing,[51] and CNN[52] routines written in-house. All computations were performed on a DEC 3000 AXP Model 500 workstation. Methods used to develop QSPRs can be described by four steps: (1) structure entry and modeling,

10.1021/ci9900130 CCC: $18.00 © 1999 American Chemical Society
Published on Web 04/10/1999

Table 1. Compounds Used in This Study

no.	compound	obsd log(k_r), M^{-1} s^{-1}	CNN model[a]	CNN model[b]	ref[c]	no.	compound	obsd log(k_r), M^{-1} s^{-1}	CNN model[a]	CNN model[b]	ref[c]
	Alkenes and Isolated Double Bonds										
1	ethene	4.519	4.665	4.395[e]	27, 29, 30, 41, 45	39	1,1-diphenylethene	6.146	5.680	6.057	21, 24, 28
2	propene	4.279	4.414	4.342	29, 33, 40, 41, 45	40	dibenzfulvene	7.230	6.396[d]	6.580	33
3	1-butene	4.362	4.411[e]	4.304	29, 30, 33, 40	41	cyclopentene	3.699	3.571	3.418	43
4	3-methyl-1-butene	4.301	4.407	4.264	29	42	cyclohexene	2.886	3.702	3.446	43
5	1-pentene	4.322	4.412	4.304	29, 33	43	cycloheptene	3.556	3.824[e]	3.705	43
6	1-heptene	4.342	4.414	4.310	29, 33	44	cyclooctene	3.763	3.760	3.633	43
7	1-decene	4.279	4.415	4.339	29, 33	45	bicyclo[2.2.1]heptene	4.633	3.699	4.405	43
8	1-hexadecene	4.342	4.416	4.510	29, 33	46	(*E*)-2-butene	3.771	3.538	3.419[d]	22, 26, 29, 30, 33
9	allyl acetate	4.857	3.595	4.700	26, 35	47	(*Z*)-2-butene	3.462	3.476	3.384[d]	22, 26, 29, 30, 33
10	ethyl vinyl ether	3.839	4.399	4.426	35	48	(*Z*)-di-*tert*-butylethene	3.204	3.370	3.474	28
11	vinyl acetate	4.462	4.547	4.849	21, 24, 35	49	(*Z*)-*β*-methylstyrene	4.531	4.707	4.038[e]	37
12	methyl acrylate	5.944	6.175	5.599[e]	44	50	(*E*)-*β*-methylstyrene	4.886	4.866	4.072	33, 37
13	methyl vinyl ketone	6.204	6.174	6.325	44	51	indene	4.914	5.064	4.042	37
14	acrylonitrile	6.176	5.980[d]	6.455	24,44	52	(*Z*)-stilbene	4.398	4.521	4.225	28
15	styrene	5.833	5.598	5.747	21, 24, 35, 33, 41	53	(*E*)-stilbene	4.954	5.017[d]	4.293	21, 23, 24, 28
16	1,4-diisopropenylbenzene	6.255	6.351	6.743	37	54	diethyl fumarate	6.230	6.282	5.296	28, 35
17	2,4,6-trimethylstyrene	4.949	5.568[e]	5.691	37	55	fumaronitrile	6.230	6.233	6.415	28, 35
18	2-chlorostyrene	5.934	5.925	6.012	37	56	diethyl maleate	5.462	5.989	5.099[e]	21, 24, 28, 35
19	3-chlorostyrene	5.949	5.917	6.003	37	57	maleic anhydride	6.568	6.482	6.775	28, 35
20	4-chlorostyrene	5.944	5.978	6.078	37	58	chloromaleic anhydride	6.763	6.509	6.640	28
21	2,5-dichlorostyrene	6.000	6.171	5.398	37	59	dichloromaleic anhydride	5.623	6.528[e]	6.194	28
22	1-vinylnaphthalene	5.845	6.039	6.378	37	60	maleonitrile	6.230	6.518	5.717	28, 35
23	1-vinylanthracene	6.079	6.362	6.333	37	61	methyl crotonate	4.763	5.483[e]	4.388	44
24	9-vinylanthracene	5.580	6.418	6.438[e]	37	62	2-methyl-2-butene	3.681	3.639	3.455	33
25	2-vinylpyridine	6.000	6.444[e]	5.799[e]	35, 37	63	1-cyanocyclopentene	5.279	4.660	5.103	44
26	4-vinylpyridine	6.079	6.317[d]	5.967[e]	35, 37	64	*β*,*β*-dimethylstyrene	4.114	4.749	4.004[e]	36
27	2-vinylthiophene	6.255	6.107	6.315[d]	37	65	*β*,*β*-dimethyl methylacrylate	4.000	4.437[e]	4.244	44
28	vinyl bromide	5.398	5.269	5.311[d]	46	66	*β*,*β*-dimethylacrylonitrile	4.301	4.425	5.121	44
29	vinyl chloride	5.279	4.909	5.035	46	67	9-ethylidenefluorene	6.447	5.627	5.586	37
30	vinyl fluoride	4.230	4.321[e]	4.927[e]	46, 47	68	9-isopropylidenfluorene	5.415	5.657	5.819[d]	37
31	2-methylpropene	4.491	4.475	4.403	29, 30, 33, 41	69	triphenylethene	4.602	4.587	4.167[d]	21, 24
32	methylenecyclobutane	4.568	4.472	4.510	42	70	trifluoroethene	4.690	5.760[d]	4.884	46, 47
33	1-phenylacetic acid vinyl ester	3.898	4.263	4.005	26	71	α,*β*,*β*-trimethylstyrene	4.230	4.367	4.046	37
34	α-methylstyrene	5.903	5.589[d]	5.758	35, 41	72	tetrafluoroethene	5.519	6.003	5.495	27, 46, 47
35	methyl methacrylate	6.079	6.316	5.342	24, 35, 44	73	1,4-pentadiene	4.708	4.545	4.795[d]	33
36	methacrylonitrile	6.255	6.105	6.361	44	74	1,5-hexadiene	4.763	4.512	4.609[d]	33
37	1,1-dichloroethene	5.944	5.677	5.332	46	75	2,5-dimethyl-1,5-hexadiene	4.820	4.914	5.381	33
38	1,1-difluoroethene	4.322	4.017	5.332	46, 47	76	bicyclo[2.2.1]heptadiene	4.996	4.340	5.090	43
	Conjugated Multiple Bonds										
77	1,3-butadiene	6.230	6.033	6.439	33, 35, 41, 45	86	1,2-dimethylene-3-methylcyclopentane	6.748	6.318[d]	6.396	42
78	isoprene	6.255	6.099	6.425[d]	33, 35, 41	87	1,4-diphenyl-1,3-butadiene	5.519	4.923[d]	4.410	33
79	chloroprene	6.813	6.514	6.195	33	88	2,5-dimethyl-2,4-hexadiene	4.362	5.112	4.577[d]	33
80	2,3-dimethyl-1,3-butadiene	6.279	6.213[d]	6.414	33, 35	89	1,1,4,4-tetraphenyl-1,3-butadiene	4.716	4.652	4.499	33
81	(*Z*)-1,3-pentadiene	6.041	5.633	5.635	33	90	cyclopentadiene	5.362	4.909[d]	4.911	43
82	(*E*)-1,3-pentadiene	5.857	5.638[e]	5.633	33	91	1,3-cyclohexadiene	5.756	5.126	5.017	43
83	4-methyl-1,3-pentadiene	5.934	5.667	5.620	33	92	2,4-hexadiene	5.176	5.034	4.915	33
84	1,2-dimethylenecyclobutane	6.613	6.232[d]	6.323[e]	42	93	cycloheptatriene	5.114	5.226	5.378	43
85	1,2-dimethylenecyclohexane	5.940	6.171	6.120	42	94	1,3,5,7-cyclooctatetraene	4.845	5.069	5.334	43
	Alkynes										
95	ethyne	4.398	4.424[d]		31	99	phenylethyne	5.230	4.446[d]		31
96	propyne	3.968	3.746[e]		31	100	2-butyne	3.079	2.384		31
97	1-pentyne	4.079	3.485[d]		31	101	diphenylethyne	4.000	4.597		31
98	1-hexyne	4.176	3.489		31						
	Cumulated Multiple Bonds										
102	allene	4.176	4.735		33	105	2,3-pentadiene	4.079	3.995		33
103	1,2-butadiene	4.114	4.881		33	106	tetraphenylallene	4.653	4.675		33
104	1,2-pentadiene	4.230	4.937[d]		33	107	vinylacetylene	6.279	5.446		44
	Benzenes										
108	benzene	2.398	2.884	3.366[d]	3, 23, 24	115	benzonitrile	3.477	3.667[e]	4.132[e]	32
109	toluene	2.613	2.820	3.398[d]	26	116	bromobenzene	2.954	3.175	3.271	32
110	biphenyl	3.114	4.173	4.112	3, 23	117	chlorobenzene	3.021	3.137	3.129	32
111	anisol	2.212	2.808	2.643	32	118	fluorobenzene	2.740	2.851	3.097	32
112	ethyl benzoate	3.114	3.924[e]	3.527	32	119	1,3-dichlorobenzene	3.491	3.291	3.156	32
113	acetophenone	2.778	3.841	3.912	32	120	1,4-dichlorobenzene	3.462	3.372	3.238	32
114	benzophenone	3.556	3.440	3.860	3, 23						

Table 1. (Continued)

no.	compound	obsd log(k_r), M^{-1} s^{-1}	CNN model[a]	CNN model[b]	ref[c]	no.	compound	obsd log(k_r), M^{-1} s^{-1}	CNN model[a]	CNN model[b]	ref[c]
	Condensed Aromatic Compounds										
121	naphthalene	3.908	3.937	4.188	3, 23, 38	142	2-fluoronaphthalene	4.041	3.733[e]	3.944	38
122	1-methylnaphthalene	3.845	3.902	4.148[d]	38	143	1-iodonaphthalene	4.204	4.397[d]	4.066	38
123	2-methylnaphthalene	4.041	3.846	4.116	38	144	2-iodonaphthalene	4.415	4.306	4.003[d]	38
124	1-ethylnaphthalene	3.839	3.862	4.085	38	145	1,5-dimethylnaphthalene	3.699	3.870	4.113	38
125	2-ethylnaphthalene	3.881	3.837	4.076	38	146	2,3-dimethylnaphthalene	3.954	3.801	4.059	38
126	1-methoxynaphthalene	3.881	3.708	3.258	38	147	2,6-dimethylnaphthalene	4.079	3.760	4.042	38
127	2-methoxynaphthalene	3.806	3.753	3.453	38	148	1,4-dichloronaphthalene	4.255	4.458	3.997	38
128	1-dimethylaminonaphthalene	3.477	3.819	3.505	38	149	acenaphthene	3.602	3.737	4.319	38
129	1-methyl naphthoate	4.230	4.343	4.020[e]	38	150	anthracene	5.519	4.989	4.957	3, 23, 36, 39
130	2-methyl naphthoate	4.544	4.013	3.890	38	151	1-methylanthracene	5.491	4.957	5.225	36
131	1-ethyl naphthoate	4.279	4.819[e]	3.948[e]	38	152	2-methylanthracene	5.491	4.921[e]	5.194	36
132	1-acetonaphthone	4.230	4.397	4.249	38	153	9-methylanthracene	5.204	4.980	5.221	36
133	2-acetonaphthone	4.580	3.990	4.160	38	154	2,6-dimethylanthracene	5.491	4.861	5.390	36
134	1-naphthonitrile	4.415	4.571	4.466	38	155	9,10-dimethylanthracene	4.716	4.970	5.457	36
135	2-naphthonitrile	4.591	4.402	4.382	38	156	1,3,5,7-tetramethylanthracene	5.491	4.775	5.157	39
136	1-naphthyl isocyanate	4.176	4.842	4.836	38	157	1,4,5,8-tetramethylanthracene	5.342	4.837	5.262[e]	39
137	1-bromonaphthalene	3.845	4.357	4.042	38	158	2,3,6,7-tetramethylanthracene	5.477	4.780[e]	5.127	39
138	2-bromonaphthalene	4.255	4.303	4.013	38	159	phenanthrene	3.944	4.328[d]	4.334	3, 23, 25, 36
139	1-chloronaphthalene	4.041	4.282	3.931[d]	38	160	2-methylphenanthrene	4.041	4.225	4.386	36
140	2-chloronaphthalene	4.279	4.260	3.919[d]	38	161	3-methylphenanthrene	4.041	4.245	4.424	36
141	1-fluoronaphthalene	3.954	3.827	3.911[e]	38	162	9,10-dimethylphenanthrene	3.716	4.239	4.493	36
	Quinones										
163	1,4-benzoquinone	6.716	6.517	6.707	22, 24	175	chloranil	5.000	4.739	5.231	22, 24
164	2-methylbenzoquinone	6.556	6.538	6.670	22, 24	176	1,2-naphthoquinone	6.079	6.544	6.223	22, 24
165	2-methoxybenzoquinone	6.431	6.533	6.895	24	177	1,4-naphthoquinone	6.447	6.511[e]	6.285	34
166	2-chlorobenzoquinone	6.949	6.543	6.961	22, 24	178	2-methyl-1,4-naphthoquinone	6.079	6.358	6.136	22, 24
167	2,3-dimethylbenzoquinone	6.431	6.542	6.534	34	179	2-chloro-1,4-naphthoquinone	6.748	6.342	6.309	34
168	2,5-dimethylbenzoquinone	6.431	6.545	6.536	24, 34	180	2,3-dimethyl-1,4-naphthoquinone	5.279	5.178	5.465	22, 24
169	2,6-dimethylbenzoquinone	6.491	6.543	6.539	34	181	2,7-dimethyl-1,4-naphthoquinone	6.146	5.955	5.821	22, 24
170	2,6-dimethoxybenzoquinone	5.944	6.550[e]	6.534[e]	34	182	2,3-dichloro-1,4-naphthoquinone	4.491	5.184	5.195	22, 24
171	2,3-dichlorobenzoquinone	6.531	6.524	6.916[d]	34	183	6,7-dichloro-1,4-naphthoquinone	6.398	6.442	5.502	34
172	2,5-dichlorobenzoquinone	7.114	6.504	6.951[e]	34	184	2-*tert*-butylanthraquinone	4.491	4.259	4.518	22, 24
173	2,6-dichlorobenzoquinone	7.114	6.508[d]	6.953	24	185	phenanthraquinone	5.380	5.369	6.127	22, 24
174	duroquinone	4.531	4.893[e]	4.776	22, 24						
	Heterocycles										
186	acridine	5.204	5.187[d]		23, 24, 39	189	1-azaanthracene	5.672	4.980		39
187	quinoline	4.079	4.017		3, 23, 39	190	4,5-dimethylacridine	5.146	5.212		39
188	phenazine	4.964	5.431		39	191	1,4,5,8-tetramethylacridine	5.079	5.092		39

[a] Type 3 model for data set 1 using CNN for feature selection and model formation. [b] Type 2 model for data set 2 using linear feature selection and CNN for model formation. [c] References refer to original work. See ref 5 for conversion to absolute rate constants. [d] Member of cross-validation set. [e] Member of external prediction set.

(2) descriptor generation and initial reduction, (3) linear model formation and validation, and (4) nonlinear model formation and validation.

Structure Entry and Modeling. Compounds were sketched, and initial three-dimensional modeling was performed using HyperChem (Hypercube, Inc., Waterloo, ON). This produced connection tables which contained atom types and bonding information. The three-dimensional structure of each compound was refined with the semiempirical molecular orbital program MOPAC[53] with the PM3 Hamiltonian.[54]

Descriptor Generation and Initial Reduction. The calculated structural descriptors used in this study fall into four classes: topological, geometric, electronic, and hybrid or combination descriptors. Topological descriptors are generated from a simple two-dimensional sketch, and they encode features such as atom types, molecular connectivity, and branching.[55−58] Geometric descriptors require an energy minimized three-dimensional structure. Examples of such descriptors are length-to-breadth ratio, moment of inertia, and solvent-accessible surface areas and volumes.[59,60] The electronic descriptors provide information such as partial atomic charges and polarizability.[61−63] Combination descriptors represent hybrids of the first three classes. One example of hybrid descriptors are charged partial surface area (CPSA) descriptors,[64] which represent combinations of partial atomic charge information (electronic) and solvent accessible surface area (geometric). A total of 206 descriptors were calculated, of which 122 were topological, 28 were geometric, 8 were electronic, and the remaining 48 were combination descriptors.

Objective feature selection was used to reduce the number of available descriptors. To accomplish this, descriptors calculated for the training set compounds were examined. Any descriptor with identical values for at least 90% of the observations was removed. Pairwise correlations were also investigated. One of any two descriptors with $r > 0.93$ was removed from the pool. This left reduced pools of 73 descriptors for data set 1 and 79 descriptors for data set 2.

Linear Model Formation. The reduced pool of descriptors was screened using evolutionary optimization procedures (genetic algorithm[50] and simulated annealing[51]). Subsets of descriptors were examined to see which could successfully map k_r based on linear regression. These type 1 linear models

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

descriptor	coefficient	error estimate	explanation
NDB	0.659	0.049	no. of double bonds
1SP2	0.630	0.093	sp^2 hybridized carbon bonded to one other carbon
MDE	0.105	0.017	distance-edge term between quaternary carbons
PND	-1.53×10^{-2}	3.7×10^{-3}	superpendentic index
MOMI	-1.21×10^{-3}	2.7×10^{-4}	third major moment of inertia
PNSA	7.33×10^{-3}	1.21×10^{-3}	partial negative surface area
constant	3.02	0.17	*Y*-intercept

were evaluated based on correlation coefficient (*R*), rms error, and descriptor *T*-values. Furthermore, variance inflation factors (VIF $= [1-R^2]^{-1}$) were calculated by regressing each descriptor against all others (excluding the dependent variable). If VIF was less than 10 for all model variables, the model was deemed free of multicollinearities.

Model size was determined by searching for the smallest subset of descriptors with the best correlation coefficient and lowest rms error. Additionally, prediction set rms error was examined to ensure predictive ability of the model. The descriptor subset identified as forming the most predictive, statistically valid linear model was saved for consideration in generating nonlinear models.

Nonlinear Model Formation. Descriptors identified during linear model formation were used to build a three layer, fully connected, feed-forward CNN. These are called type 2 models. All networks consisted of an input layer with the number of neurons determined by the linear model, a hidden layer with a varying number of neurons, and an output layer with one neuron to produce k_r values. Network architectures were varied in order to find the point at which an additional hidden layer neuron did not reduce rms error of the training and cross-validation sets. The prediction set was only used to monitor predictive ability of the model.

A final model, type 3, was developed using CNN with nonlinear feature selection. The original reduced pool of descriptors (73 for data set 1, 79 for data set 2) was screened using a genetic algorithm with a neural network as the fitness evaluator. Descriptor subsets identified in this manner were used to fully train a neural network. Network architecture was determined as described previously. Predictive ability was again ensured with the external prediction set.

RESULTS AND DISCUSSION

Data Set 1. Many descriptor subsets and subset sizes were considered for regression model formation. Descriptor subset size was determined by starting with a three-descriptor model and finding the point at which subsequent addition of a descriptor did not improve rms error. Table 2 defines the best six-descriptor type 1 model identified. All *T*-values were greater than 4, and no VIF exceeded 10. This model had $R = 0.793$ and training set rms error $= 0.677$ log units. Prediction set rms error was 0.630 log units, which demonstrates the ability of the model to generalize. The magnitude of *R* and rms error values will be discussed more fully below.

Of the six descriptors in the linear model, four were topological, one geometric and one combination descriptor. Pairwise correlations for the six descriptors ranged from 0.060 to 0.594 with an average value of 0.278. NDB encodes the number of double bonds, and 1SP2 encodes the number of carbon atoms bonded to two hydrogens and one carbon with sp^2 hybridization. These descriptors probably encode information about attack sites for the radical on the substrate and ability to stabilize the product radical through resonance. MDE represents a molecular edge descriptor[65] probably describing quaternary carbons or lack of attacking sites for the radical. PND denotes the superpendentic index,[66] which encodes degree of branching and perhaps captures steric information. The CPSA descriptor, PNSA, describes partial negative surface area and probably encodes electronic information regarding energetics of products, reactants, and transition states.

As stated above, the type 1 model has R = 0.793 and a training set rms error = 0.677 log units. It should be noted that experimental error varies from compound to compound, and the only definite statement made about the data as a whole is that statistical error is below 20%.[5] Some of this is due to experimental error, and some due to the conversion of methyl affinities measured by Swarc.[5] With this in mind, results of the linear model do indicate a successful QSPR. Additionally, randomization of the dependent variable produced a six-descriptor regression model with R = 0.331 and rms errors of 1.04 log units for the training set and 1.13 log units for the prediction set. Clearly, the type 1 model described in Table 2, is superior to this and demonstrates that the results achieved were not due to chance.

The six descriptors forming the linear model were then used to generate a type 2 CNN model. The input layer consisted of the six descriptors in Table 2 and the output layer contained one neuron for k_r. The number of hidden layer neurons was varied to determine the optimum network architecture. The number of hidden layer neurons was kept as small as possible without compromising network performance.

Network training was accomplished using the BFGS (Broyden−Fletcher−Goldfarb−Shanno) method.[67] Training was directed by optimization of weights and biases to minimize training set error. Steps were taken to avoid overtraining the network. The ratio of observations to adjustable parameters was kept above 2.0. Additionally, the cross-validation set rms error was used to terminate training. When the cross-validation set rms error no longer decreased, training was terminated, as this point suggests the network is beginning to memorize idiosyncrasies of the training set members. The final network had a 6-4-1 architecture (33 adjustable parameters).

In general, prediction errors will be lower when CNNs are used as opposed to regression. This is most likely due to the increased number of adjustable parameters and the nonlinear capabilities of CNNs. In this study, training set rms error for the type 2 nonlinear model was reduced 17.4% to 0.559 log units. Prediction set rms error was 0.574 log units, a decrease of 8.9%. Cross-validation set rms error for this network was 0.688 log units.

Finally, a type 3 model was generated by selecting subsets of descriptors from the reduced pool using a genetic algorithm with a CNN as the fitness evaluator. The final nonlinear type 3 model chosen also had a 6-4-1 architecture, and the descriptors selected are shown in Table 3. For these six descriptors, pairwise correlations ranged from 0.136 to 0.763 with an average pairwise correlation of 0.449. The two topological descriptors (NDB and 1SP2) chosen were

Table 3. The Type 3 Six-Descriptor Model Developed Nonlinear Feature Selection for Data Set 1

descriptor	explanation
NDB	no. of double bonds
1SP2	sp^2 hybridized carbon bonded to one other carbon
LUMO	energy of lowest unoccupied molecular orbital
PNSA	atomic charge weighted partial negative surface area
SAAA	(Σ surface area of acceptor atoms)/ (total molecular surface area)
SCAA	(Σ (surface area of acceptor atoms)*(charge on acceptor atom))/(total molecular surface area)
constant	*Y*-intercept

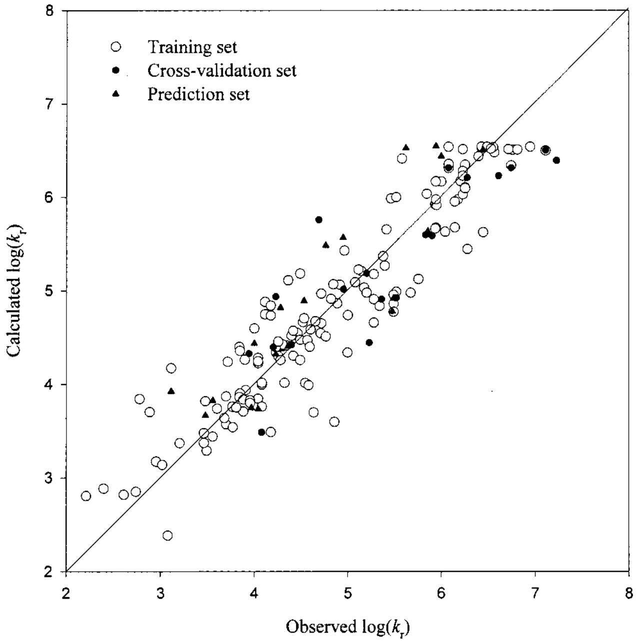

Figure 1. Plot of calculated vs observed log(k_r) of the training, cross-validation, and prediction set compounds for data set 1. Rate constants were calculated using the CNN described in Table 3.

also seen in the linear model (Table 2). The electronic descriptor, LUMO, represents the energy of the lowest unoccupied molecular orbital. All three hybrid descriptors (WPNSA, SAAA, SCAA) probably work to encode information about charge distribution and electron density.

Figure 1 shows a plot of calculated vs observed log(k_r) values for the training, cross-validation, and prediction sets using this model. The rms errors are training set = 0.381 log units (31.9% improvement), cross-validation set = 0.511 log units (25.6% improvement), and prediction set = 0.496 log units (13.5% improvement), where percent improvement is relative to the type 2 CNN model. Note that the rms errors for the training, cross-validation, and prediction sets do vary somewhat for this model. However, predictive ability is clearly improved in going from linear feature selection and regression (type 1) to linear feature selection and CNN (type 2) and finally to nonlinear feature selection and CNN (type 3).

Data Set 2. As stated in the Experimental Section, models were built on a subset of the original 191 compounds. Allenes, alkynes, and heterocycles were removed due to the small number of compounds forming each class (see Table 1). These class sizes may not be large enough to ensure adequate representation in selecting descriptors. As a group, these compounds seemed to have unusually high rms errors for some models. For example, the type 2 model produced rms error for these 19 compounds of 0.804 log units. The remaining 172 compounds have rms error = 0.543 log units. Similar observations were made for some of the other models investigated. Therefore, models were generated using the 172 remaining compounds.

Table 4. Seven-Descriptor Type 1 Linear Regression Model Selected by Evolutionary Optimization Techniques for Data Set 2

descriptor	coefficient	error estimate	explanation
ALLP	3.69×10^{-3}	5.1×10^{-4}	total no. of paths in structure up to length 45
WTPT	−0.155	0.028	sum of all path weights starting from heteroatoms
1SP2	1.16	0.09	sp^2 hybridized carbon bonded to one other carbon
PND	-1.72×10^{-2}	3.0×10^{-3}	superpendentic index
MCB	−0.235	0.021	number of multiple carbon−carbon bonds
GEOM	1.81	0.42	third major moment
LUMO	−1.56	0.10	energy of lowest unoccupied molecular orbital
constant	5.59	0.15	*Y*-intercept

Descriptor reduction based on the 155 training set compounds resulted in a reduced pool of 79 descriptors. Linear feature selection allowed evaluation of regression models. The seven-descriptor model described in Table 4 was chosen. Pairwise correlations ranged from 0.073 to 0.845 with an average of 0.286. Topological descriptors 1SP2 and PND and electronic descriptor LUMO are again present. Three new topological descriptors (ALLP, WTPT, MCB) were also selected. ALLP describes all paths in a molecule, and WTPT denotes weighted paths from heteroatoms. Both descriptors probably encode steric information. MCB is the number of multiple carbon−carbon bonds (double, triple, or aromatic), which probably encodes information on attack sites and potential for resonance stabilization of the product radical. The geometric descriptor GEOM represents the third major moment for the molecule.

This type 1 regression model produced a training set rms error = 0.555 log units, an 18.0% improvement over the regression model for data set 1. The prediction set rms error was 0.543 log units, representing a 13.7% decrease in error. The regression model still has error higher than the errors obtained for data set 1 using nonlinear feature selection and CNNs.

After removal of 17 compounds from the training set to form a cross-validation set, the seven descriptors used to form the regression model were used to build a type 2 CNN model. A 7-3-1 architecture (28 adjustable parameters) was selected. Figure 2 shows a calculated vs observed log(k_r) plot for the training, cross-validation, and prediction sets. For the training set, rms error was 0.424 log units. This is about 11.4% higher than the error associated with the type 3 model for data set 1. However, the cross-validation set rms error was 0.399 log units, which is a 22.0% improvement versus the cross-validation rms error in Figure 1. Furthermore, prediction set rms error reduced to 0.409 log units, a 17.6% decrease relative to Figure 1.

Overall, the model described in Figure 2 is superior to the model described in Figure 1. The rms error values are more consistent across the training, cross-validation, and prediction sets. The ratio of adjustable parameters to

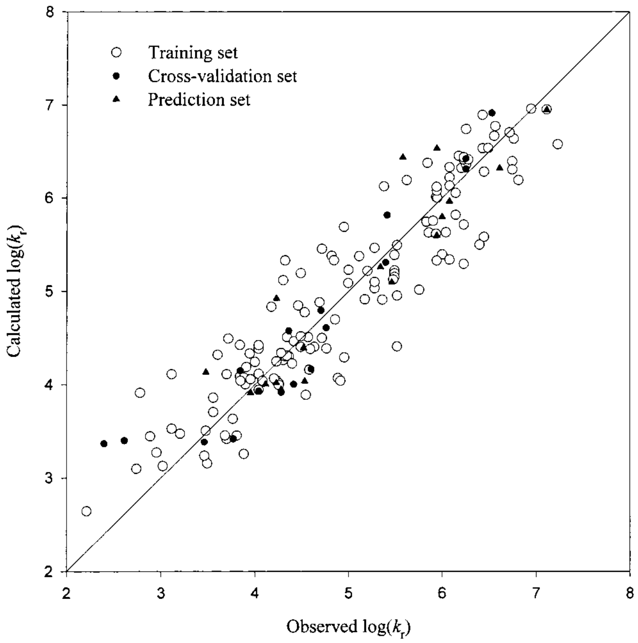

Figure 2. Plot of calculated vs observed log(k_r) of the training, cross-validation, and prediction set compounds for data set 2. Rate constants were calculated using the CNN described in Table 4.

observations for this model (4.92) is somewhat better than for the previous model (4.64). No models employing nonlinear feature selection could be found for data set 2 that better the model developed using linear feature selection and a CNN (Figure 2).

CONCLUSIONS

In this study, six-descriptor models were built for data set 1 using multiple linear regression and CNNs. The type 3 nonlinear model with the lowest rms error values resulted from using nonlinear feature selection and a 6-4-1 CNN. Resulting rms errors were training set = 0.381 log units, cross-validation set = 0.511 log units, and prediction set = 0.496 log units. Generation of QSPRs suggests that descriptors calculated encode steric hindrance, polar effects, radical stability, and other factors influencing reaction rates.

The models generated for data set 2 demonstrated that predictive ability can be improved by reducing the variability of the data (training set rms error = 0.424 log units, cross-validation set rms error = 0.399 log units, prediction set rms error = 0.409 log units). Although it would be ideal to build one model for all compounds, this cannot be accomplished without proper amounts of training data for each compound class. More data for allenes, alkynes, and heterocycles would demonstrate whether subsetting would be necessary to maintain an rms error of approximately 0.4 log units, which appears to be on the order of experimental error.

ACKNOWLEDGMENT

The authors are grateful to Dr. K. Héberger for assistance with data collection and helpful discussions.

REFERENCES AND NOTES

(1) Giese, B. Formation of CC Bonds by Addition of Free Radicals to Alkenes *Angew. Chem., Int. Ed. Engl.* **1983**, *22*, 753−764.

(2) Slanina, J.; Warneck, P.; Bazhin, N. M.; Akimoto, H.; Kieskamp, W. M. Assessment of Uncertainties in the Projected Concentrations of Methane in the Atmosphere. *Pure Appl. Chem.* **1994**, *66*, 137−200.

(3) Levy, M.; Szwarc, M. Methyl Affinities of Aromatic Compounds *J. Chem. Phys.* **1954**, *22*, 1621−1622.

(4) Zytowski, T.; Fischer, H. Absolute Rate Constants for the Addition of Methyl Radicals to Alkenes in Solution: New Evidence for Polar Interactions. *J. Am. Chem. Soc.* **1996**, *118*, 437−439.

(5) Zytowski, T.; Fischer, H. Absolute Rate Constants and Arrhenius Parameters for the Addition of the Methyl Radical to Unsaturated Compounds: The Methyl Affinities Revisited. *J. Am. Chem. Soc.* **1997**, *119*, 12869−12878.

(6) Houk, K. N.; Paddon-Row: M. N.; Spellmeyer, D. C.; Rondan, N. G.; Nagase, S. Theoretical Transition Structures for Radical Additions to Alkenes. *J. Org. Chem.* **1986**, *51*, 2874−2879.

(7) Arnaud, R.; Subra, R.; Barone, V.; Lelj, F.; Olivella, S.; Solé, A.; Russo, N. Ab initio Mechanistic Studies of Radical Reactions. Directive Effects in the Addition of Methyl Radical to Unsymmetrical Fluoroethanes. *J. Chem. Soc., Perkin Trans. 2* **1986**, 1517−1524.

(8) Fueno, T.; Kamachi, M. Ab Initio SCF Study of the Addition of the Methyl Radical to Vinyl Compounds. *Macromolecules* **1988**, *21*, 908−912.

(9) Gonzalez, C.; Sosa, C.; Schlegel, H. B. Ab Initio Study of the Addition Reaction of the Methyl Radical to Ethylene and Formaldehyde. *J. Phys. Chem.* **1989**, *93*, 2435−2440.

(10) Wong, M. W.; Pross, A.; Random, L. Addition of Methyl Radical to Substituted Alkenes: A Theoretical Study of the Reaction Mechanism. *Isr. J. Chem.* **1993**, *33*, 415−425.

(11) Davis, T. P.; Rogers, S. C. Ab Initio Molecular Orbital Calculations on the Transition State for the Addition of a Methyl Radical to Vinyl Monomers. *Macromol. Theory Simul.* **1994**, *3*, 905−913.

(12) Alfrey, T., Jr.; Price, C. C. Relative Reactivities in Vinyl Copolymerization. *J. Polym. Sci.* **1947**, *2*, 101−106.

(13) Colthup, N. B. Molecular Orbital Study of the Q-e Scheme in Free Radical Copolymerization. *J. Polym. Sci.* **1982**, *20*, 3167−3179.

(14) Rogers, S. C.; Mackrodt, W. C.; Davis, T. P. Ab Initio Molecular Orbital Calculations on the Q-e Scheme for Predicting Reactivity in Free-Radical Copolymerization. *Polym.* **1994**, *35*, 1258−1267.

(15) Bamford, C. H.; Jenkins, A. D.; Johnston, R. Patterns of Free Radical Reactivity. *Trans. Faraday Soc.* **1959**, *55*, 418−433.

(16) Ito, O.; Matsuda, M. A New Dual-Parameter for Reactivities of Vinyl Monomers toward Free-Radicals. *J. Polym. Sci., Part A: Polym. Chem.* **1990**, *28*, 1947−1963.

(17) Atkinson, R. A Structure−Activity Relationship for the Estimation of Rate Constants for the Gas-Phase Reactions of OH Radicals with Organic Compounds. *Int. J. Chem. Kinet.* **1987**, *19*, 799−828.

(18) Klamt, A. Estimation of Gas-Phase Hydroxyl Radical Rate Constants of Organic Compounds from Molecular Orbital Calculations. *Chemosphere* **1993**, *26*, 1273−1289.

(19) Klamt, A. Estimation of Gas-Phase Hydroxyl Radical Rate Constants of Oxygenated Compounds Based on Molecular Orbital Calculations *Chemosphere* **1996**, *32*, 717−726.

(20) Medven, Ž.; Güsten, H.; Sabljić, A. Comparative QSAR Study on Hydroxyl Radical Reactivity with Unsaturated Hydrocarbons: PLS Versus MLR. *J. Chemom.* **1996**, *10*, 135.

(21) Leavitt, F.; Levy, M.; Szwarc, M.; Stannett, V. Methyl Affinities of Vinyl Monomers. Part I. Styrene and Phenylated Ethylenes. *J. Am. Chem. Soc.* **1955**, *77*, 5493−5497.

(22) Rembaum, A.; Szwarc, M. Methyl Affinities of Quinones. *J. Am. Chem. Soc.* **1955**, *77*, 4468−4472.

(23) Levy, M.; Szwarc, M. Reactivities of Aromatic Hydrocarbons toward Methyl Radicals. *J. Am. Chem. Soc.* **1955**, *77*, 1949−1955.

(24) Szwarc, M. Reactions of Methyl Radicals and Their Applications to Polymer Chemistry. *J. Polym. Sci.* **1955**, *16*, 367−382.

(25) Szwarc, M. Singlet−Triplet Excitation Energy of Aromatic Compounds and Their Reactivities. *J. Chem. Phys.* **1955**, *23*, 204−206.

(26) Buckley, R. P.; Leavitt, F.; Szwarc, M. Reactions of Methyl Radicals with Substrates Acting as Hydrogen Donors and as Methyl Radical Acceptors. *J. Am. Chem. Soc.* **1956**, *78*, 5557−5560.

(27) Buckely, R. P.; Szwarc, M. Methyl Affinities of Ethylene, Tetrafluoroethylene and Tetrachloroethylene. *J. Am. Chem. Soc.* **1956**, *78*, 5696−5697.

(28) Bader, A. R.; Buckley, R. P.; Leavitt, F.; Szwarc, M. Addition of Methyl Radical to cis and trans Isomers. *J. Am. Chem. Soc.* **1957**, *79*, 5621−5625.

(29) Buckley, R. P.; Szwarc, M. The Addition of Methyl Radicals to Ethylene, Propylene, the Butenes and Higher 1-Olefines. *Proc. R. Soc.* **1957**, *A240*, 396−407.

(30) Buckley, R. P.; Rembaum, A.; Szwarc, M. Methyl Affinities of Vinyl Monomers. Ethylene and Its Homologues. *J. Polym. Sci.* **1957**, *24*, 135−137.

(31) Gazith, M.; Szwarc, M. Addition of Methyl Radicals to Acetylenic Compounds. *J. Am. Chem. Soc.* **1957**, *79*, 3339−3343.

(32) Heilman, W. J.; Rembaum, A.; Szwarc, M. Addition of Methyl Radicals to Substituted Benzenes. *J. Chem. Soc.* **1957**, 1127−1130.

(33) Rajbenbach, A.; Szwarc, M. Addition of Methyl Radicals to Isolated, Conjugated and Cumulated Dienes. *Proc. R. Soc. (London)* **1957**, *A251*, 394−406.

(34) Buckley, R. P.; Rembaum, A.; Szwarc, M. Addition of Methyl Radicals to Quinones. Part II. The Reaction Centre. *J. Chem. Soc.* **1958**, 3442−3445.

(35) Leavitt, F.; Stannett, V.; Szwarc, M. Relative Selectivity of Polystyryl Radicals. *J. Polym. Sci.* **1958**, *31*, 193−195.

(36) Binks, J. H.; Szwarc, M. Effect of Conjugation, Hyperconjugation, and Steric Hindrance on Methyl Affinities. *J. Chem. Phys.* **1959**, *30*, 1494−1501.

(37) Carrock, F.; Szwarc, M. Methyl Affinities of Substituted Styrenes, their Homologues and Analogues. *J. Am. Chem. Soc.* **1959**, *81*, 4138−4144.

(38) Gresser, J.; Binks, J. H.; Szwarc, M. Effect of Substituents on Methyl Affinity of Naphthalene Derivatives. *J. Am. Chem. Soc.* **1959**, *81*, 5004.

(39) Binks, J. H.; Gresser, J.; Szwarc, M. The Configuration of the Transition State in Methyl-Radical Additions. *J. Chem. Soc.* **1960**, 3944−3947.

(40) Steel, C.; Szwarc, M. Methyl Affinities Determined by Photolysis of Azomethane. *J. Chem. Phys.* **1960**, *33*, 1677−1680.

(41) Feld, M.; Szwarc, M. The Entropy of Activation of Addition of Methyl Radicals to Unsaturated Compounds Possessing the Same Reaction Center. *J. Am. Chem. Soc.* **1960**, *82*, 3791−3792.

(42) Gresser, J.; Rajbenbach, A.; Szwarc, M. Relation Between Methyl Affinities and Conformation of the Conjugated Dienes. *J. Am. Chem. Soc.* **1960**, *82*, 5820−5822.

(43) Gresser, J.; Rajbenbach, A.; Szwarc, M. Methyl Affinities of Some Cyclic Olefins and Polyenes. *J. Am. Chem. Soc.* **1961**, *83*, 3005−3008.

(44) Herk, L.; Stefani, A.; Szwarc, M. Methyl Affinities of Some Compounds Related to Acrylates and Acrylonitriles. Reactivities of Conjugated Systems Involving Atoms Other Than Carbon. *J. Am. Chem. Soc.* **1961**, *83*, 3008−3011.

(45) Feld, M.; Stefani, A. P.; Szwarc, M. The Secondary Deuterium Effect in CH_3 and CF_3 Addition Reactions. *J. Am. Chem. Soc.* **1962**, *84*, 4451−4456.

(46) Stefani, A. P.; Todd, H. E. Kinetics of Addition of Cyclopropyl Radicals to Olefins. *J. Am. Chem. Soc.* **1970**, *93*, 2982−2986.

(47) Sass, V. P.; Serov, S. I.; Sokolov, S. V. Reactivity of Fluorine-Substituted Ethylenes. Addition of a Methyl Radical. *Zh. Org. Khim. (Engl. Ed.)* **1977**, *13*, 2298−2300.

(48) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.

(49) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979.

(50) 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.

(51) 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.

(52) 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.

(53) Stewart, J. P. P. *Mopac 6.0*, *Quantum Chemistry Program Exchange*; Indiana University, Bloomingtom, IN, Program 455.

(54) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

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

(56) Randiæ, 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.

(57) Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

(58) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd.: Hertfordshire, England, 1986.

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

(60) 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.

(61) Miller, K. J.; Savchik, J. A. A New Empirical Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.

(62) 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.

(63) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.

(64) 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.

(65) Mitchell, B. E. Ph.D. Thesis, The Pennsyvania State University.

(66) 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.

(67) 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.

CI9900130