Neural Network–Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds

Andrei A. Gakh,[*] Elena G. Gakh, Bobby G. Sumpter, and Donald W. Noid

Chemistry Division,[†] Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6197

Received October 29, 1993[⊗]

A new computational scheme is developed to predict physical properties of organic compounds on the basis of their molecular structure. The method uses graph theory to encode the structural information which is the numerical input for a neural network. Calculated results for a series of saturated hydrocarbons demonstrate average accuracies of 1–2% with maximum deviations of 12–14%.

I. INTRODUCTION

Determination of the physical properties of organic molecules based on their structures is in the main stream of computational chemistry. Despite the amount of literature available on the subject of quantitative structure/property relationships (QSPR) and large achievements of the last decade, existing computational schemes need further improvement.

Implementation of computational artificial intelligence methods could help provide a breakthrough in this direction. In particular, the relatively newly developed and now commercially available multilayer computer neural networks seem to be valuable tools for the prediction of properties (as is shown herein).

Only recently have neural networks been extensively applied to chemistry problems.[1a–c] The most successful chemical applications of neural networks were found for the prediction of a property at given conditions based on the knowledge of this property under other conditions (extrapolation).[2a–c] It is clear that the success of the neural network methods for these applications is due to the continuous character of the input/output functional relationships which are favorable for the neural network.

The use of neural networks for QSPR is somewhat different from other computational methods and has both advantages and disadvantages. The primary advantage is that the self-adjusting (back-propagation) algorithm can generalize "knowledge" obtained through the training process without the need of theoretical formulas or postulated models. The disadvantages is the necessity of preprocessing of the structural information into a form acceptable for neural networks. The numerical inputs must represent the structural features of molecules as distinctly as possible without any additional complexity, thus allowing the neural network to elucidate QSPR with the maximum efficiency. The neural network performance and the accuracy of the results are strongly dependent on the way the structural data preprocessing was performed.

One of the simplest approaches for employing neural networks in QSPR problems can be based on the use of structural formulas of compounds as a source of information for generating inputs. A topological algorithm for the preprocessing of structural formulas has been chosen in order to prepare the numerical input for a back-propagation-type neural network. Although graph theory (mainly in the form of topological indexes) was actively used for QSPR since the mid 1940s, its combination with artificial intelligence computational principles provides a new powerful research tool with many extended capabilities.

II. METHODOLOGY

A. Neural Networks and Neural Computing. The theory and general practice of artificial neural network applications is already well documented.[1a–c,2b,3,4] We used these concepts to create our own back-propagation-type neural network simulator for a mainframe computer environment.[5] The neural network simulator (computer program written in C) uses the learning procedure of back-propagation and can be summarized as follows:

(1) Initialize the node connection weights w_{ij} to some small random values.

(2) Input some data V^m_i and corresponding output values V^T_i, where m is the layer number, i is the node number, and T represents the target or desired output state.

(3) Propagate the initial signal forward through the network using

$$\mathrm{NET}^m_j = \sum w^m_{ij} V^{m-1}_i + \gamma_j \quad \text{where} \quad V^m_j = F(\mathrm{NET}^m_j) \qquad (1)$$

where w_{ij} is the connection weights between nodes i and j, V^{m-1}_i is the signal from node i and layer $m = 1$, γ is the threshold or bias value of the node, and F is a transfer function and is generally taken as the sigmoid function.

$$F(\mathrm{NET}^m_j) = 1/[1 + \exp(-2\beta \mathrm{NET}^m_j)] \qquad (2)$$

We have chosen to write the sigmoid function in this way in order to demonstrate its similarity to a commonly known function in physics called the Fermi function. The parameter β can be used to adjust the sigmoid to vary between a linear (small values) and a step function (large values). The feed-forward propagation (eq 1 is continued for each i and m until the final outputs V°_i have been calculated.

(4) Compute the deltas (δ) for the output layer, defined as

$$\delta^\circ_i = -\partial E/\partial \mathrm{NET}^\circ_i = -(\partial E/\partial V^\circ_i)\,(\partial V^\circ_i/\partial \mathrm{NET}^\circ_i) = F'(\mathrm{NET}^\circ_i)\,[V^T_i - V^\circ_i] \qquad (3)$$

where the error function E is taken to be proportional to the square of the difference between the actual and desired (target) output and $F'(\mathrm{NET}^\circ_i)$ is the derivative of the transfer function

[†] Research sponsored by the Division of Materials Sciences, Office of Basic Energy Sciences, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc., and by a CRADA, under DOE contract, No. ORNL92-0118, between Martin Marietta Energy Systems and Hoechst Celanese Corp.

[⊗] Abstract published in *Advance ACS Abstracts,* April 1, 1994.

0095-2338/94/1634-0832$04.50/0 © 1994 American Chemical Society

Table 1. List of 109 Hydrocarbons and Their Properties Used as a Learning Set[a]

hydrocarbon	C_n	heat capacity, at 300 K	boiling point, °C	density kg/m^3, at 25 °C	refractive index, at 25 °C	Gibbs energy ΔG, at 300 K	enthalpy 300 K
3-methylpentane	C_6	140.88	63.28	659.76	1.3739	–2.12	26.32
		145.87	68.05	676.22	1.3827	–2.16	26.75
2,2-dimethylbutane	C_6	142.26	49.74	644.46	1.3660	–7.42	25.40
		143.27	55.00	664.33	1.3695	–7.67	25.61
2,3-dimethylbutane	C_6	140.21	57.99	657.02	1.3723	–1.77	24.77
		143.21	62.25	679.98	1.3778	–1.76	25.12
3-methylhexane	C_7	164.50	91.85	682.88	1.3861	6.60	30.71
		167.87	93.55	691.13	1.3919	6.70	27.73
3-ethylpentane	C_7	166.80	93.48	693.92	1.3911	12.70	31.71
		169.66	94.53	690.53	1.3971	12.47	31.84
2,2-dimethylpentane	C_7	167.70	79.17	669.48	1.3800	2.10	29.50
		167.23	81.86	673.16	1.3818	2.21	29.76
2,3-dimethylpentane	C_7	161.80	89.75	690.81	1.3895	7.60	28.62
		165.94	91.47	705.69	1.3943	7.73	29.04
2,4-dimethylpentane	C_7	171.70	80.47	668.23	1.3788	4.90	29.58
		170.54	84.60	681.49	1.3840	5.15	29.95
3,3-dimethylpentane	C_7	166.70	86.04	689.16	1.3884	4.80	29.33
		167.05	89.21	709.59	1.3920	4.95	29.51
2,2,3-trimethylbutane	C_7	164.20	80.86	685.64	1.3869	6.30	28.28
		162.54	84.93	708.19	1.3858	6.20	28.18
n-octane	C_8	188.70	125.68	698.54	1.3951	17.67	38.12
		188.36	120.06	695.07	1.3947	16.97	37.64
2-methylheptane	C_8	188.20	117.65	693.87	1.3926	13.37	35.82
		188.83	113.99	688.22	1.3998	13.23	35.80
3-methylheptane	C_8	186.82	118.93	701.73	1.3961	13.79	35.31
		189.10	117.23	705.35	1.3990	13.91	35.92
2,4-dimethylhexane	C_8	193.35	109.43	696.17	1.3929	13.07	33.76
		192.04	112.26	710.57	1.3966	13.71	33.08
2,5-dimethylhexane	C_8	186.52	109.11	689.37	1.3900	11.40	33.39
		188.45	109.95	693.10	1.3937	11.74	33.92
3,3-dimethylhexane	C_8	191.96	111.97	705.95	1.3978	15.13	33.43
		190.67	113.22	718.04	1.3995	15.56	33.65
3,4-dimethylhexane	C_8	182.72	117.73	715.15	1.4018	18.43	32.47
		190.30	118.41	729.88	1.4056	18.73	33.40
3-ethyl-2-methylpentane	C_8	193.05	115.66	715.20	1.4017	20.68	34.31
		192.95	115.33	713.40	1.4049	21.00	34.45
3-ethyl-3-methylpentane	C_8	189.07	118.27	723.54	1.4055	24.36	33.26
		190.73	118.57	732.95	1.4084	23.83	33.48
2,2,3-trimethylpentane	C_8	186.77	109.84	712.03	1.4007	19.45	32.13
		186.89	109.18	726.95	1.4000	19.35	32.14
2,2,4-trimethylpentane	C_8	189.45	99.24	687.84	1.3890	15.70	32.55
		189.56	102.17	700.51	1.3899	16.42	32.73
2,3,3-trimethylpentane	C_8	188.20	114.77	722.30	1.4052	20.04	32.17
		187.52	113.69	747.95	1.4054	19.91	32.11
2,3,4-trimethylpentane	C_8	192.72	113.47	715.09	1.4020	20.76	32.55
		190.95	113.49	733.91	1.4047	20.99	32.60
2,2,3,3-tetramethylbutane	C_8	188.28	106.29	821.70		24.04	31.84
		190.33	118.41	729.88	1.4057	18.73	33.40
2-methyloctane	C_9	210.90	143.28	709.60	1.4008	21.60	40.42
		209.18	139.38	709.25	1.3983	21.37	40.25
3-methyloctane	C_9	209.70	144.23	716.70	1.4040	22.00	39.92
		210.45	141.39	719.53	1.4033	22.08	39.81
3-ethylheptane	C_9	213.00	143.20	722.50	1.4070	26.40	40.71
		213.09	139.97	717.22	1.4072	25.98	40.36
4-ethylheptane	C_9	214.30	141.20	722.30	1.4067	26.80	40.50
		213.18	139.50	718.46	1.4065	26.26	40.26
2,2-dimethylheptane	C_9	212.40	132.82	706.60	1.3995	19.50	38.83
		209.83	128.78	695.52	1.3979	19.38	38.70
2,3-dimethylheptane	C_9	207.70	140.50	722.00	1.4064	23.50	37.82
		210.21	131.75	725.13	1.4077	23.49	37.99
2,4-dimethylheptane	C_9	217.10	133.20	711.50	1.4011	20.80	38.16
		215.37	134.98	718.12	1.4039	21.93	38.39
2,5-dimethylheptane	C_9	208.20	136.00	713.60	1.4015	18.20	37.53
		212.02	136.51	728.32	1.4060	22.23	37.73
2,6-dimethylheptane	C_9	210.40	135.22	704.50	1.3985	19.80	37.99
		211.50	132.62	698.01	1.4027	19.82	38.38
3,3-dimethylheptane	C_9	214.00	137.02	721.60	1.4063	22.00	38.20
		212.03	136.51	718.05	1.4050	22.46	37.73
3,4-dimethylheptane	C_9	206.80	140.40	727.50	1.4091	24.90	37.02
		211.77	140.93	737.57	1.4113	25.70	37.42
3,5-dimethylheptane	C_9	214.60	135.70	716.60	1.4046	22.00	38.07
		214.34	138.99	734.24	1.4078	23.02	38.28
3-ethyl-3-methylhexane	C_8	214.10	140.60	736.00	1.4134	30.50	37.36
		213.95	141.26	741.64	1.4135	30.42	37.62
4-ethyl-2-methylhexane	C_9	219.70	133.80	724.20	1.4054	24.50	39.25
		215.51	135.95	721.72	1.4071	25.03	39.26
2,2,4-trimethylhexane	C_9	210.70	129.91	711.80	1.4010	23.60	36.61
		211.48	130.97	724.91	1.4016	24.61	36.84
2,2,5-trimethylhexane	C_9	209.10	124.09	703.20	1.3973	15.30	36.36
		209.67	125.97	706.11	1.3987	16.36	36.87
2,3,3-trimethylhexane	C_9	213.30	137.69	733.50	1.4119	29.40	36.28
		211.53	136.53	746.63	1.4112	29.36	36.33
2,3,4-trimethylhexane	C_9	214.00	138.96	735.10	1.4120	28.60	36.86
		213.87	140.14	752.64	1.4147	29.49	36.91
2,3,5-trimethylhexane	C_9	212.50	131.36	717.90	1.4037	22.20	36.02
		213.13	134.44	734.01	1.4074	23.54	36.52

Table 1 (Continued)

hydrocarbon	C_n	heat capacity, at 300 K	boiling point, °C	density kg/m^3, at 25 °C	refractive index, at 25 °C	Gibbs energy ΔG, at 300 K	enthalpy 300 K
2,4,4-trimethylhexane	C_9	213.50	130.66	720.05	1.4052	26.60	36.44
		211.80	133.91	738.12	1.4070	27.75	37.97
3,3,4-trimethylhexane	C_9	210.50	149.45	741.40	1.4154	31.40	35.98
		211.78	144.12	758.20	1.4159	31.64	36.06
3,3-diethylpentane	C_9	217.86	146.19	749.92	1.4184	43.30	38.37
		217.29	143.91	740.16	1.4196	41.08	38.43
3-ethyl-2,2-dimethylpentane	C_9	205.00	133.84	731.00	1.4101	37.50	36.15
		210.68	132.63	729.73	1.4090	36.66	36.36
3-ethyl-2,3-dimethylpentane	C_9	213.40	144.70	750.80	1.4197	36.80	36.61
		211.76	140.73	753.02	1.4164	35.17	36.54
2,2,3,3-tetramethylpentane	C_9	213.34	140.29	752.97	1.4214	39.00	35.86
		208.55	139.24	765.86	1.4167	36.54	35.18
2,2,3,4-tetramethylpentane	C_9	208.50	133.03	735.22	1.4125	36.70	35.06
		209.90	131.91	754.19	1.4116	36.15	34.90
2,3,3,4-tetramethylpentane	C_9	219.50	141.56	751.11	1.4200	39.70	36.23
		212.61	137.80	770.63	1.4188	37.99	35.59
3-ethyloctane	C_{10}	235.80	166.50	735.90	1.4136	34.90	45.31
		233.91	165.87	732.85	1.4091	34.47	44.80
4-ethyloctane	C_{10}	236.50	163.64	734.30	1.4131	33.40	45.10
		234.66	163.49	731.51	1.4093	33.14	44.61
2,2-dimethyloctane	C_{10}	235.10	156.90	720.80	1.4060	27.70	43.43
		229.85	155.64	712.85	1.4000	27.53	43.27
2,4-dimethyloctane	C_{10}	239.40	155.90	722.60	1.4069	28.80	42.76
		235.93	160.96	731.30	1.4060	30.22	42.96
2,5-dimethyloctane	C_{10}	231.80	158.50	726.40	1.4089	26.90	41.92
		233.54	162.15	729.78	1.4089	28.38	38.34
3,4-dimethyloctane	C_{10}	229.30	163.40	741.80	1.4159	33.00	41.80
		233.59	165.50	747.67	1.4155	33.78	42.06
3,5-dimethyloctane	C_{10}	238.30	159.40	732.90	1.4115	29.10	42.47
		237.41	164.20	743.31	1.4124	31.00	42.67
3,6-dimethyloctane	C_{10}	229.60	160.80	732.90	1.4115	28.90	41.63
		233.88	164.33	741.88	1.4136	29.99	42.09
4,4-dimethyloctane	C_{10}	239.30	157.50	731.20	1.4122	31.90	42.30
		236.20	160.52	736.34	1.4094	33.06	42.45
4,5-dimethyloctane	C_{10}	230.10	162.13	743.20	1.4167	35.30	41.51
		234.55	165.59	747.48	1.4161	36.22	41.92
4-*n*-propylheptane	C_{10}	237.70	157.50	732.10	1.4113	38.20	44.85
		235.93	159.39	728.90	1.4087	36.45	44.36
4-isopropylheptane	C_{10}	239.20	158.90	735.40	1.4132	37.90	43.10
		236.66	159.97	733.79	1.4120	37.07	43.08
2-methyl-3-ethylheptane	C_{10}	238.50	161.20	739.80	1.4151	35.70	43.30
		237.17	160.07	732.01	1.4149	35.55	43.13
2-methyl-4-ethylheptane	C_{10}	243.40	156.20	732.20	1.4114	31.60	43.64
		238.21	158.00	730.23	1.4117	32.09	43.51
3-methyl-4-ethylheptane	C_{10}	236.20	162.20	746.60	1.4183	36.90	42.47
		235.98	163.32	746.03	1.4159	36.91	49.47
3-methyl-5-ethylheptane	C_{10}	240.90	158.20	736.80	1.4141	33.10	43.35
		237.71	161.43	741.89	1.4127	37.76	43.15
2,2,3-trimethylheptane	C_{10}	232.50	157.60	738.50	1.4145	34.80	41.30
		231.92	154.64	739.20	1.4126	33.94	41.13
2,2,4-trimethylheptane	C_{10}	234.70	148.30	725.70	1.4092	31.90	41.05
		235.82	151.55	728.66	1.4200	33.17	41.23
2,2,5-trimethylheptane	C_{10}	230.50	150.80	728.10	1.4101	23.80	40.50
		233.27	152.69	726.59	1.4127	25.32	40.98
2,2,6-trimethylheptane	C_{10}	234.80	148.93	723.80	1.4178	24.20	40.96
		234.34	148.12	704.22	1.4225	24.72	41.43
2,3,3-trimethylheptane	C_{10}	235.10	160.20	748.80	1.4202	37.30	41.00
		233.61	159.05	752.76	1.4190	36.55	41.56
2,3,4-trimethylheptane	C_{10}	237.60	159.90	748.50	1.4195	37.20	40.96
		237.75	162.54	756.94	1.4220	38.20	41.10
2,3,5-trimethylheptane	C_{10}	233.90	160.70	745.10	1.4169	30.30	40.12
		236.49	163.12	753.11	1.4214	32.13	41.26
2,3,6-trimethylheptane	C_{10}	228.50	156.00	734.70	1.4131	28.50	39.75
		234.12	157.63	734.64	1.4193	29.47	40.80
2,4,4-trimethylheptane	C_{10}	238.90	151.00	734.60	1.4143	35.90	40.54
		239.14	156.17	741.29	1.4162	37.84	40.30
2,4,5-trimethylheptane	C_{10}	234.10	156.50	737.30	1.4160	36.10	39.98
		237.67	161.56	750.62	1.4207	33.45	40.99
2,4,6-trimethylheptane	C_{10}	246.30	147.60	719.00	1.4071	28.40	41.13
		241.85	154.43	729.28	1.4139	30.60	40.76
2,5,5-trimethylheptane	C_{10}	234.20	152.80	736.20	1.4149	25.80	40.54
		235.30	156.86	740.90	1.4181	27.65	41.93
3,3,5-trimethylheptane	C_{10}	234.10	155.68	739.00	1.4170	34.10	40.46
		234.29	159.21	755.17	1.4132	35.18	40.60
3,4,4-trimethylheptane	C_{10}	235.60	161.10	753.50	1.4235	40.30	40.08
		236.89	163.18	761.80	1.4235	41.17	40.56
3,4,5-trimethylheptane	C_{10}	235.10	162.50	751.90	1.4229	39.70	41.14
		237.88	167.09	768.10	1.4261	40.96	40.43
2-methyl-3-isopropylhexane	C_{10}	231.80	166.70	743.60	1.4172	46.80	40.46
		236.20	163.25	744.58	1.4180	45.67	41.40
3,3-diethylhexane	C_{10}	242.50	166.30	757.50	1.4235	51.00	42.43
		239.74	166.40	749.53	1.4222	48.15	41.10
3,4-diethylhexane	C_{10}	246.90	163.90	747.20	1.4167	45.40	43.68
		239.93	164.56	745.63	1.4136	43.66	42.51
2,2-dimethyl-3-ethylhexane	C_{10}	227.70	156.10	744.70	1.4174	44.40	40.50
		232.65	155.69	738.45	1.4148	42.90	43.02

Table 1 (Continued)

hydrocarbon	C_n	heat capacity, at 300 K	boiling point, °C	density kg/m^3, at 25 °C	refractive index, at 25 °C	Gibbs energy ΔG, at 300 K	enthalpy 300 K
2,2-dimethyl-4-ethylhexane	C_{10}	236.10	147.00	730.20	1.4107	36.10	41.92
		235.17	154.15	732.73	1.4101	36.29	40.94
2,3-dimethyl-3-ethylhexane	C_{10}	238.20	163.70	759.90	1.4247	45.00	40.71
		237.33	164.61	763.52	1.4242	44.48	40.79
2,3-dimethyl-4-ethylhexane	C_{10}	243.00	160.90	751.60	1.4203	42.10	42.43
		239.78	163.92	757.86	1.4221	42.54	41.88
2,4-dimethyl-4-ethylhexane	C_{10}	235.00	160.10	751.40	1.4202	42.30	40.29
		237.34	163.60	757.83	1.4220	42.68	42.23
3,3-dimethyl-4-ethylhexane	C_{10}	228.20	162.90	759.80	1.4246	50.00	39.92
		233.16	162.66	757.31	1.4204	47.18	41.25
3,4-dimethyl-4-ethylhexane	C_{10}	235.50	162.10	759.60	1.4244	47.60	40.42
		236.67	166.18	770.73	1.4232	46.55	40.16
2,2,3,3-tetramethylhexane	C_{10}	238.20	160.31	760.89	1.4260	48.70	40.00
		232.79	156.52	768.16	1.4213	45.72	39.77
2,2,3,4-tetramethylhexane	C_{10}	229.40	158.80	751.30	1.4193	46.10	39.25
		233.36	159.62	769.20	1.4194	45.69	39.76
2,2,3,5-tetramethylhexane	C_{10}	235.80	148.40	733.60	1.4119	32.10	39.54
		235.06	152.28	750.84	1.4140	34.04	39.56
2,2,4,5-tetramethylhexane	C_{10}	229.20	147.88	731.61	1.4132	32.80	38.83
		234.07	152.93	747.34	1.4165	35.00	40.44
2,2,5,5-tetramethylhexane	C_{10}	229.80	137.46	714.80	1.4055	21.10	39.37
		231.70	142.80	719.15	1.4074	23.27	39.55
2,3,3,4-tetramethylhexane	C_{10}	241.50	164.59	765.60	1.4298	49.10	40.04
		237.31	164.69	782.22	1.4298	48.29	39.14
2,3,5-tetramethylhexane	C_{10}	234.00	153.10	744.90	1.4196	41.20	39.16
		234.78	157.28	751.36	1.4179	40.82	40.35
2,3,4,4-tetramethylhexane	C_{10}	231.80	161.60	758.60	1.4267	49.20	38.87
		234.20	160.47	756.50	1.4230	47.03	39.52
2,3,4,5-tetramethylhexane	C_{10}	243.10	156.20	745.60	1.4204	42.70	40.71
		239.79	162.13	771.87	1.4249	44.04	39.19
3,3,4,4-tetramethylhexane	C_{10}	238.00	170.00	778.90	1.4368	58.90	39.87
		235.23	166.33	787.52	1.4345	54.33	39.88
2,4-dimethyl-3-isopropylpentane	C_{10}	234.50	157.04	754.57	1.4246	64.30	39.25
		240.48	159.95	759.85	1.4378	61.72	39.99
2-methyl-3,3-diethylpentane	C_{10}	224.80	169.70	775.50	1.4320	60.30	39.25
		236.41	167.41	761.25	1.4322	57.41	39.81
2,2,3-trimethyl-3-ethylpentane	C_{10}	223.70	169.50	778.00	1.4397	66.60	38.41
		230.60	161.84	769.86	1.4338	68.34	39.04
2,2,4-trimethyl-3-ethylpentane	C_{10}	227.30	155.30	753.10	1.4199	59.00	38.87
		235.89	155.84	754.09	1.4198	57.00	38.76
2,3,4-trimethyl-3-ethylpentane	C_{10}	229.00	169.44	773.50	1.4310	61.00	38.58
		230.27	162.47	783.89	1.4243	59.62	38.32
2,2,3,3,4-pentamethylpentane	C_{10}	234.30	166.05	776.75	1.4341	68.80	38.62
		231.78	159.06	791.57	1.4299	68.44	37.63
2,2,3,4,4-pentamethylpentane	C_{10}	234.20	159.29	763.61	1.4281	63.80	38.81
		233.73	154.22	776.05	1.4261	58.82	37.96

[a] The top numbers are the experimental values, and the bottom are those calculated by a neutral network after finishing the optimization procedure.

with respect to the activation NET_i. For the sigmoid function, the derivative is

$$F'(\mathrm{NET}^\circ{}_i) = \partial F(\mathrm{NET}^\circ{}_i)/\partial \mathrm{NET}^\circ{}_i = 2\beta F(\mathrm{NET}^\circ{}_i)[1 - F(\mathrm{NET}^\circ{}_i)] \quad (4)$$

where $F(\mathrm{NET}^\circ{}_i)$ is given by eq 2.

(5) Compute the deltas for the preceding layers by propagating the errors backward:

$$\delta^{m-1}{}_i = F'(\mathrm{NET}^{m-1}{}_i)[\sum w_{im}{}^m \delta^m{}_j] \quad (5)$$

for all m = m, $m-1$, $m-2$, ..., until delta has been calculated for every node.

(6) Using

$$\Delta w^m{}_{ij} = \eta \delta^m{}_i V^{m-1}{}_j \quad (6)$$

update the connection weights to

$$w^{\mathrm{new}}{}_{ij} = w^{\mathrm{old}}{}_{ij} + \Delta w_{ij} \quad (7)$$

In eq 6, the parameter η is called the learning rate and controls the magnitude of the effects of the deltas on the connection weight updates.

(7) Return to step 2 and repeat for another input example. This process is continued until the network output satisfies some error criteria. There are a number of useful modifications that can be used to help speed convergence and improve generalization accuracy. For the work presented in this paper, the standard back-propagation described above with the addition of biases, short-cut connections (direct connections between the input and the output nodes), momentum, and an adaptive learning rate was all that was required to make accurate predictions. The simulator, however, offers a variety of possibilities that can be invaluable for other problems. For a more complete description, the reader is referred to ref[2b].

B. Preprocessing of the Structural Information for the Neural Network. The preprocessing, or a numerical transformation of those structural graphs[6] which most accurately represent chemical information for the neural network to make correct predictions, remains a fundamental problem.[1a,6] A numerical structural representation for many of the existing databases has usually been done using connectivity tables, which in turn are based on the structural graphs.[7a,b,8a] These tables typically contain the numerated atoms and their connections as combinations of these numbers.[8a] Unfortunately, the connectivity tables are not unique since there are in general many ways to numerate the atoms. To avoid this ambiguity, one needs to employ very sophisticated preference algorithms[8a,b] or all possible connectivity tables need to be included in any subsequent calculations. These sophisticated rules may be difficult to recognize by neural networks. On

Table 2. List of 25 Hydrocarbons and Their Properties Used as a Testing Set[a]

hydrocarbon	C_n	heat capacity, at 300 K	boiling point, °C	density kg/m^3, at 25 °C	refractive index, at 25 °C	Gibbs energy ΔG, at 300 K	enthalpy 300 K
2-methylpentane	C_6	143.01	60.27	648.52	1.3687	–4.05	26.61
		146.76	64.67	656.01	1.3766	–4.19	26.12
n-heptane	C_7	166.0	98.40	679.50	1.3851	9.50	33.56
		167.8	95.83	674.67	1.3896	9.11	33.35
2-methylhexane	C_7	165.4	90.03	674.34	1.3823	4.90	31.21
		167.2	90.65	674.46	1.3868	4.93	31.46
4-methylheptane	C_8	188.03	117.71	700.71	1.3955	17.40	35.06
		188.54	116.78	703.25	1.3987	17.39	34.96
3-ethylhexane	C_8	190.58	118.54	709.45	1.3992	18.53	36.07
		191.10	117.48	705.74	1.4018	18.35	36.07
2,2-dimethylhexane	C_8	189.33	106.84	691.11	1.3910	12.15	34.23
		187.58	106.24	687.65	1.3903	12.21	34.34
2,3-dimethylhexane	C_8	185.18	115.61	708.16	1.3988	17.20	33.05
		188.18	115.06	716.18	1.3918	17.24	32.67
4-methyloctane	C_9	210.40	142.44	716.30	1.4041	21.00	39.71
		211.39	141.15	718.92	1.4032	21.66	39.72
4,4-dimethylheptane	C_9	217.20	134.90	718.30	1.4053	25.80	37.53
		215.17	135.87	724.44	1.4056	26.60	37.79
3-ethyl-2-methylhexane	C_9	216.10	138.00	729.00	1.4091	26.40	38.70
		214.42	137.68	724.92	1.4098	26.82	38.77
3-ethyl-4-methylhexane	C_9	215.20	140.40	738.00	1.4128	29.90	38.07
		214.34	140.53	737.13	1.4118	30.05	38.17
2,2,3-trimethylhexane	C_9	209.90	133.58	725.70	1.4082	27.20	36.61
		209.24	132.11	734.10	1.4063	26.97	36.61
3-ethyl-2,4-dimethylpentane	C_9	209.00	136.73	734.10	1.4115	37.80	36.07
		214.17	136.89	736.69	1.4118	37.84	36.45
2,2,4,4-tetramethylpentane	C_9	215.77	122.29	715.61	1.4046	35.60	36.44
		213.03	122.53	722.33	1.4031	35.51	36.17
2,3-dimethyloctane	C_{10}	230.50	164.31	734.40	1.4127	32.00	42.43
		231.39	163.48	736.89	1.4107	31.81	42.52
2,6-dimethyloctane	C_{10}	231.90	160.38	723.60	1.4084	26.90	42.09
		233.12	162.07	726.71	1.4101	27.63	42.56
2,7-dimethyloctane	C_{10}	233.20	159.87	720.20	1.4062	28.20	42.58
		231.14	161.45	713.45	1.4063	28.11	43.19
3,3-dimethyloctane	C_{10}	237.10	161.20	735.10	1.4142	30.50	42.80
		234.02	160.88	737.84	1.4119	30.74	42.66
2-methyl-5-ethylheptane	C_{10}	234.30	159.70	731.80	1.4111	31.20	42.93
		234.46	159.60	735.82	1.4093	31.69	42.79
3-methyl-3-ethylheptane	C_{10}	236.20	163.80	746.30	1.4185	38.20	42.13
		236.30	164.28	749.11	1.4178	37.79	42.17
4-methyl-3-ethylheptane	C_{10}	238.60	163.00	746.80	1.4184	38.60	42.51
		238.14	163.51	741.53	1.4172	38.63	42.59
4-methyl-4-ethylheptane	C_{10}	239.20	160.80	747.20	1.4187	40.30	41.46
		237.98	163.66	748.35	1.4178	40.04	41.76
3,3,4-trimethylheptane	C_{10}	233.60	161.90	752.70	1.4236	38.70	40.46
		234.05	161.83	755.19	1.4216	38.31	40.64
2,5-dimethyl-3-ethylhexane	C_{10}	240.80	154.10	736.80	1.4232	33.60	41.34
		238.86	158.50	739.99	1.4260	35.29	40.96
2,2,4,4-tetramethylhexane	C_{10}	239.20	153.80	742.40	1.4185	44.10	40.29
		237.33	155.05	753.94	1.4188	44.65	39.61
			Average Error				
		0.87%	1.19%	0.60%	0.19%	1.36%	1.42%

[a] The top numbers are the experimental values, and the bottom are those predicted by the neural network.

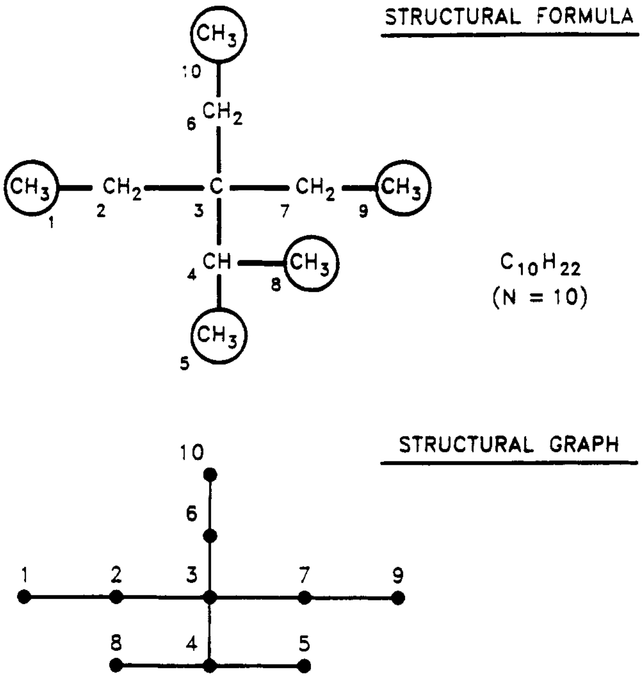
the other hand, the inclusion of all possible connectivity tables means creating a very large number of neural network inputs.

A standard approach to solve this problem is to create structural graph invariants, i.e. topological indexes. There are many known algorithms, and more are being developed.[9] Our experience indicates that there is a significant loss of structural information during the data processing, resulting in the preferential selection of topological indexes for the specific problem. In other words, some properties are best approximated with one chosen descriptor, others with different types. This is a very well-known problem in using topological indexes for QSPR.

An alternative approach involves an algorithm which generates simple, unique, and representative numerical input (structural graph invariants set), suitable for neural network operations without significant loss of structure/property relationship information. Such an algorithm was developed on the basis of Wiener's ideas on the generation of topological indexes.[10] We found then to be especially effective for our application.

We used the described back-propagation feed-forward neural network to make predictions of several physical properties for a series of saturated hydrocarbons using the set of Weiner-type descriptors. We have used hydrocarbons—a traditional test for new QSPR computational schemes—because of the availability of a large number of their properties and because they are industrially important. In addition, these compounds contain only two different atoms with the simple stoichiometric formula (C_nH_{2n+2}), making the calculations easier. To avoid difficulties, we have excluded cyclic structures from consideration (they comprise less than 30% of known saturated hydrocarbons).

The set of Weiner-type structural graph invariants were calculated by using the numbers of pathways of a given length (Figure 1).[10,11] In addition to the structural information, stoichiometric formulas (the number of carbon atoms) were



NUMERALIZATION

	PATHWAYS (n_l)	Σn_l
l = 3	5-4-8	1
l = 4		0
l = 5	1-2-3-4-5, 1-2-3-6-10, 1-2-3-7-9, 10-6-3-4-5, 1-2-3-4-8, 10-6-3-4-8, 10-6-3-7-9, 9-7-3-4-8, 9-7-3-4-5	9
l = 6		0
l = 7		0
l = 8		0

Figure 1. Sample input generation from the alkane used in our computations.

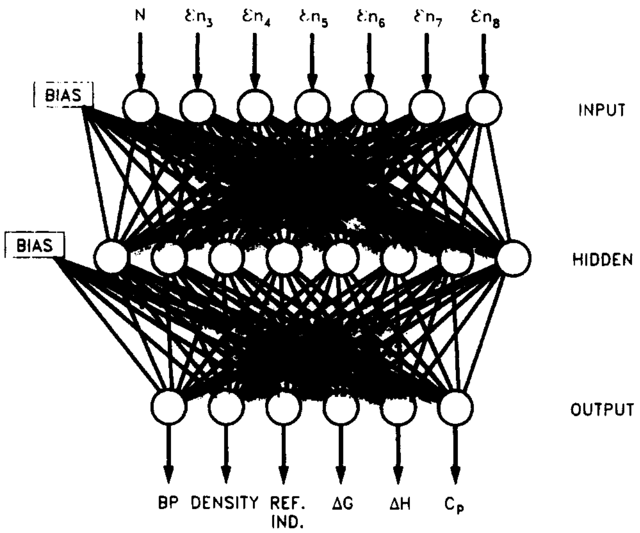


Figure 2. Diagram of the optimum neural network architecture: seven input, eight hidden, and six output nodes, plus bias nodes, arranged in three layers. Short-cut connections between the input and output nodes were also used but are not shown in this diagram.

used as independent input in the neural network computations. The intrinsic and the most important physical properties of hydrocarbons (boiling points, densities, heat capacities, standard enthalpies of formation, etc.) were used as the desired outputs. We have tried to use the most reliable sources of information to avoid experimental errors.[12]

III. RESULTS AND DISCUSSION

A neural network program[5,13] with the modifications described above has been applied to learn and predict the structure/property relationships for the set of data for 134 C_6–C_{10}. The network architecture which worked best for us (see Figure 2) consisted of two layers (one hidden and one output layer, not counting the input layer) with a topology of seven input nodes, eight hidden nodes, and six output nodes. Bias nodes and short-cut connections were also included, thus giving a total of 166 connection weights. Training of the neural network made use of momentum (0.9) and an adaptive learning rate based on the gradient. The other adjustable parameters were taken as $\beta = 0.5$, and the initial connection weight range was between –0.5 and 0.5. The final architecture was determined from a number of experiments which were aimed at finding a relatively small network capable of predicting the training data consisting of 109 examples and of making generalizations to a test set consisting of 25 examples. To ensure that the network did not overfit the traning data (memorize it), the leave-k-out method combined with cross validation was employed. This method involves taking out k examples from the training set and using them as a cross validation set (data which is used to determine the proper stopping point for the training phase: minium in the error). Depending on the size of k, many different training runs can be completed. These results are than averaged to give a representative estimate of the neural networks' performance. In addition, complexity regulation (connection weight decay) was used to sort out the topology (number of nodes in the hidden layer) of the neural network. These two precedures (leave-k-out cross validation and complexity regulation) provide a relatively high degree of confidence that the neural network topology is optimal for the given problem.

Final summarized results of the learning are presented in Table 1. These results were obtained after 300 epochs (cycles through the whole data set) of training, which took about 5 min of real time. Absence of systematic errors and a relatively low level of both average deviation (1.3–2.7%) and maximum deviation (14.6%) are unmistakable signs that the chosen neural network can correlate structural and stoichiometric parameters with the physical properties for a representative set of hydrocarbons. This trained neural network was then used as a computational tool for the prediction of the properties of a set of 25 hydrocarbons with diverse structures and molecular composition.

The results of the predictions are presented in Table 2 and in Figure 3–8. As can be seen, the neural network predicts the desired properties with an average error of less than 2% and maximum deviation of less than 12%. Average deviations of calculated data from the experimental results are presented as a histogram in Figure 9.

Only a few successful applications of neural networks for structure/property correlations are known to the authors,[1a,14a,b] and thus a cross comparison is hard to make. However, the comparison with other known computational methods based on the structural graph representation indicates that our method is among the most accurate ones.[15]

The analysis of the data obtained shows that very different physical properties (e.g., boiling points and heat capacities) of the compounds could be predicted with low average errors (0.19–1.42%). This implies that our computational model correctly represents structure/property relationships in these types of compounds and that our numerical input corresponds to structural features of the hydrocarbons responsible for their physical characteristics. It is also worth noting that the average errors in both the training and prediction sets are the same,

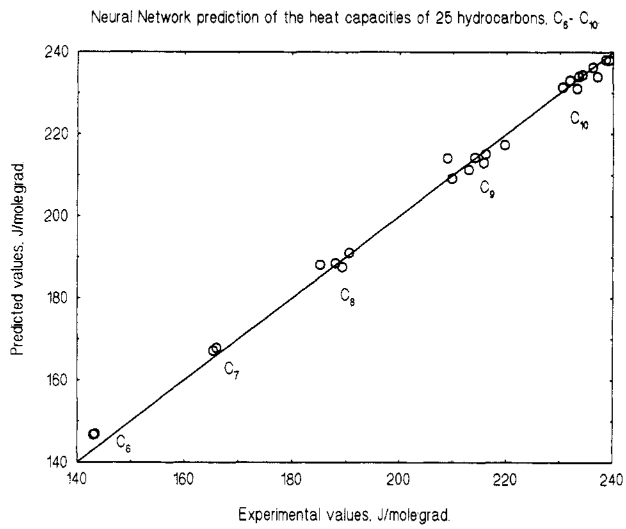


Figure 3.

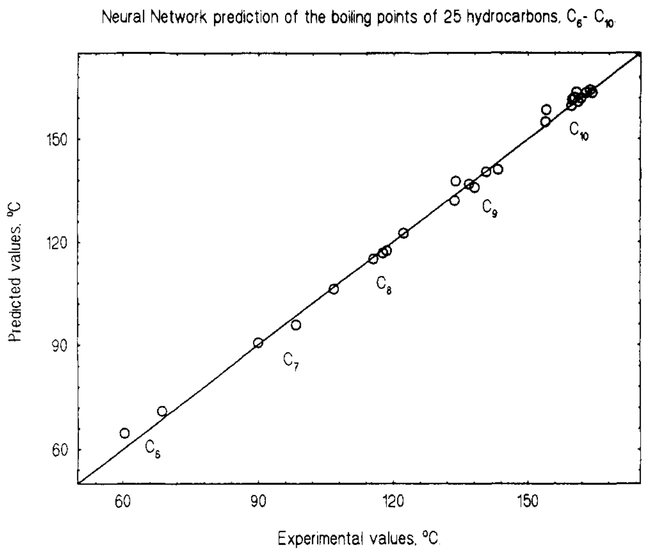


Figure 4.

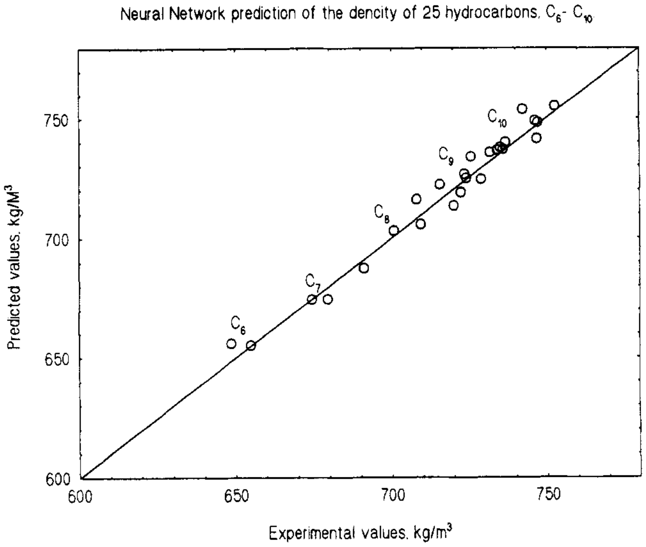


Figure 5.

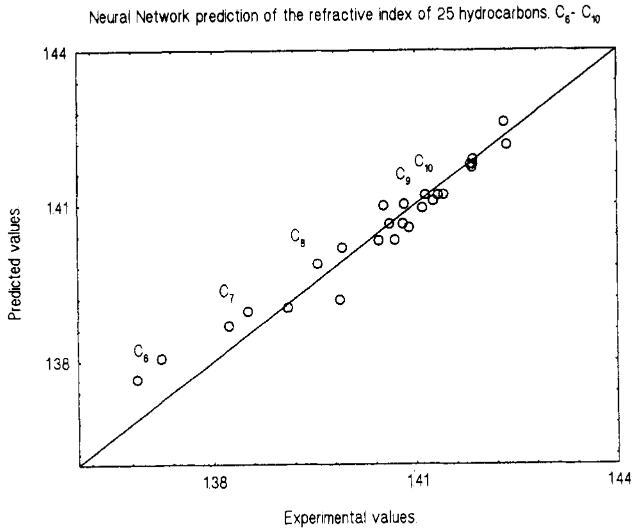


Figure 6.

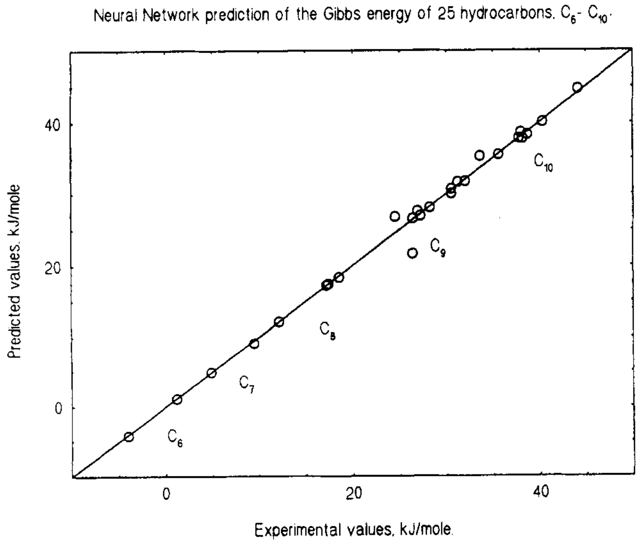


Figure 7.

which suggests that after training the neural network represents a correct "image" of the structure/physical properties in hydrocarbons.

Analysis of the results of the neural network computations revealed that the best predictions were achieved for the refractive index and density of the hydrocarbons. These parameters are not as sensitive to the specific, unique structural features of the molecules. Our computations average the structural features in the path numbers and allow effective approximation of such parameters. The same arguments can be applied to heat capacities, which were predicted with an accuracy as high as 0.9%.

The predictions of boiling points were slightly less accurate (1.2%), and those of thermodynamic functions such as Gibbs energy of formation and enthalpy were even worse (1.4–1.5%). The main reason for this is the known strong dependence of these properties on nonvalence interactions. These nonvalence interactions were not elucidated in a full range by the chosen set of descriptors. As a result, thermodynamic properties of sterically hindered and strained molecules were predicted with less accuracy than those of nonhindered ones (2–3% on average, compared to 1–2%). Addition of a parameter which could reflect such nonvalence features of molecules might be of help to improve the accuracy in these cases.

There is also a noticeable difference between the accuracies of the predictions of the properties of low (C_6) and high (C_{10}) hydrocarbons, the average error being significantly higher for hexanes (2.49%) than for decanes (0.69%). The same trend was also observed for the training set (see Table 1). We suggest, that this is mainly an effect of an inadequate training set: there are only 5 isomers of hexane compared to more than 70 structural isomers of decane.

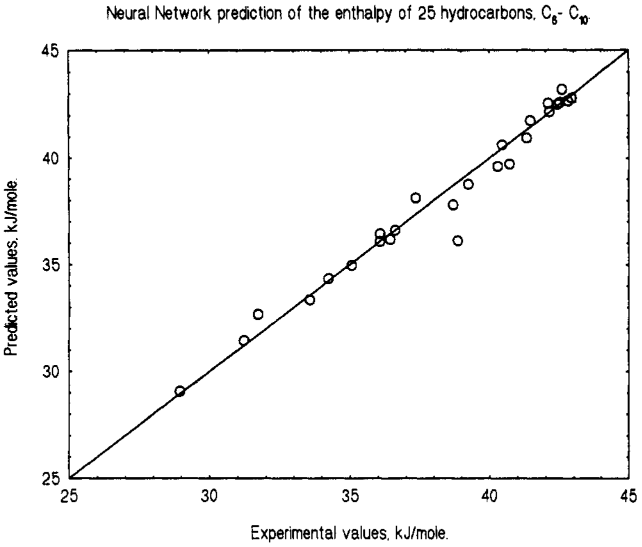


Figure 8.

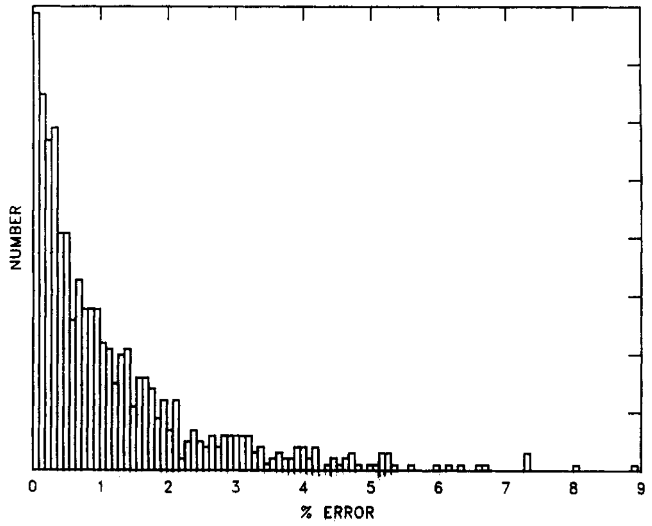


Figure 9. Histogram of the absolute percent error, abs[(experiment – network)/experiment] × 100, from the neural network calculations.

Other important information was obtained from the analysis of the neural network. Preliminary data show that the importance of the input values quickly diminishes with the increasing length of the pathways. This indicates that the structurally important information is concentrated in short-range (two to four bonds) pathways. We assume that, for more complex molecules containing heteroatoms, the length of the "important" pathways may be limited to five bonds, thus reducing the structure-related input to only four digits.

IV. CONCLUSIONS

We have presented some results on calculations of quantitative structure/property relations for saturated hydrocarbons using a neural network–graph theory approach. Very promising results were obtained with this new method, which gave up to 98% correct predictions. In addition, the overall scheme for computation is very efficient, both cost- and computer CPU-wise, thus providing a vast range for potential use in, for example, material sciences. Future studies will be focused on extending the present method to the prediction of basic and pharmacological properties of more complex organic molecules containing heteroatoms, rings, and multiple bonds.

ACKNOWLEDGMENT

A.A.G. is grateful to professor R. E. Guenther for helpful discussion and critical review of the article.

REFERENCES AND NOTES

(1) For a recent review on the applications of neural networks in chemistry see, for example: (a) Maggiora, G. M.; Elrod, D. W.; Trenary, R. G. Computational Neural Networks as Model-Free Mapping Devices. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 732–741. (b) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503–527. (c) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Anal. Chim. Acta* **1991**, *248*, 1–30.

(2) (a) Darsey, J. A.; Noid, D. W.; Wunderlich, B.; Tsoukalas, L. Neural-Net Extrapolations of Heat Capacities of Polymers to Low Temperatures. *Makromol. Chem. Rapid Commun.* **1991**, *12*, 325–330. (b) Sumpter, B. G.; Getino, C.; Noid, D. W. Neural Network Predictions of Energy Transfer in Macromolecules. *J. Phys. Chem.* **1992**, *96*, 2761–2767. (c) Noid, D. W.; Varma-Nair, M.; Wunderlich, B.; Darsey, J. A. Neural Network Inversion of the Tarasov Used for the Computation of Polymer Heat Capacities. *J. Therm. Anal.* **1991**, *37*, 2295–2300.

(3) See, for example: Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley: Redwood City, CA, 1991; Chapters 4, 5, and references cited therein.

(4) Kosko, B. *Neural Networks and Fuzzy Systems*; Prentice-Hall: New York, 1992. Cox, E. *AI Expert* June 1992, 43–47. Caudill, M.; Butler, C. Understanding Neural Networks: Computer Explorations. *Advanced Networks*; MIT Press: Cambridge, MA, 1992; Vol. 2.

(5) A program called LaMente that was developed at Oak Ridge National Laboratory under CRADA research.

(6) Elrod, D. W.; Maggiora, G. M.; Trenary, R. G. Applications of Neural Networks in Chemistry. 1. Predictions of Electrophilic Aromatic Substitution Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 477–484.

(7) (a) Lynch, M. F.; Harrison, J. M.; Town, W. G.; Ash, J. E. *Computer Handling of Chemical Structure Information*; Macdonald: London; Elsevier: New York, 1971; pp 12–35. (b) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood Ltd: Chichester, U.K., 1985; pp 128–156.

(8) (a) For a technique used in the popular CAS database see, for example: Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113. (b) For further development of Morgan numbers see, for example: Wipke, W. D.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834–4842.

(9) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological Indices for Structure-Activity Correlations. *Top. Curr. Chem.* **1983**, *114*, 21–55.

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

(11) The algorithm is based on the calculations of the shortest bond distances (pathways) between terminal carbon atoms (in our case CH_3). Then we calculate the numbers of pathways with the same length, "local sum". (As the length of pathway here, we used a number of the carbon–carbon bonds between the atoms.) For simplicity, we considered all the pathways with the same length as equal, and the pathways with a length of more than eight bonds were omitted. With these simplifications we could reduce the structural data part of inputs to only five-digit numbers. Although these simplifications have led to loss of the reverse uniqueness (we have obtained the same "local sum" numbers for some hydrocarbons C_nH_{2n+2} of the different structures with *n* more than 10), these examples are rare. A sample of an alkane structural graph representation, "local sum" numbers calculations, and final input generation (which include also *n* for a chosen C_nH_{2n+2}) is shown in Figure 1.

(12) American Petroleum Institute Research Project 44 at the National Bureau of Standards, 1947–1991, Physical and Thermodynamical Properties of Hydrocarbons.

(13) Neural network simulation computational packages are also commercially available from NeuralWare, Inc. (Penn Center West, Building IV, Pittsburgh, PA 15276–9910) both for PC and mainframe computers and from other sources.

(14) (a) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480–9483. (b) Egolf, L. M.; Jurs, P. C. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616–625 and references therein.

(15) Razinger, M.; Chretien, J. R.; Dubois, J. E. Structural Selectivity of Topological Indexes in Alkane Series. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 23–27.