

Prediction of the Clearing Temperatures of a Series of Liquid Crystals from Molecular Structure

Stephen R. Johnson and Peter C. Jurs*

152 Davey Lab, Department of Chemistry, Penn State University, University Park, Pennsylvania 16802

Received September 24, 1998. Revised Manuscript Received February 4, 1999

A quantitative structure–property relationship (QSPR) investigation was performed to develop a mathematical link between molecular structure and the clearing temperature of a series of structurally related liquid crystals. Molecular structures were encoded by a series of numerical descriptors encoding information regarding size, shape, and the ability to participate in intermolecular interactions. A genetic algorithm feature selection routine was utilized to select high-quality subsets of these descriptors for use in computational neural network models. A successful 10-descriptor model was developed using 318 compounds with a root-mean-square error of 5.4 K for the clearing temperature for the compounds in an external prediction set not used in model development.

Introduction

The use of liquid crystals in any particular application, such as electrooptical displays, requires that several physical properties be targeted for optimization. Among these properties are viscosity, elastic constants, refractive indices, and dielectric constants.¹ However, the most basic requirement is that the liquid crystalline phase must exist in an appropriate temperature range for the desired application.¹

For quite some time, there has been an interest in correlating molecular structure with liquid crystal transition temperatures. Correlations of the anisotropic molecular polarizability ($\Delta\alpha$) with the nematic to isotropic clearing temperatures (T_{NI}) were used to explain the well-known odd–even effect present in homologous series.² Maier and Saupe³ introduced a mean field approximation for the prediction of T_{NI} upon which much of the later theory has been based. Knaack et al.⁴ developed a group contribution method for the prediction of T_{NI} for structures containing two phenyl rings. This method was recently refined and extended to include a wider diversity of substituents and linker groups but was still limited to two aromatic rings.⁵ Recently, computational neural networks were used in conjunction with a modified group contribution approach for the prediction of smectic⁶ and nematic⁷ transition temperatures.

Quantitative structure–property relationship (QSPR) methodology has been used quite extensively in the literature to predict many physicochemical properties, such as boiling points,^{8,9} chromatographic retention,^{10,11} aqueous¹² and supercritical CO₂¹³ solubilities, vapor pressure,¹⁴ and polymer glass transition temperatures.¹⁵ Essentially identical methods are used for the prediction of biological activities, including acute toxicity,¹⁶ human intestinal absorption,¹⁷ and nonlethal mammalian endpoints.¹⁸ Here we present the development of QSPRs using computational neural networks (CNNs) to predict the clearing temperatures of a series of liquid crystals. Numerical descriptors are calculated to encode features of the chemical structures, which can then be linked to the target temperatures using CNNs. This approach may enable one to draw some conclusions regarding the structural characteristics required for thermostability.

Experimental Section

Data Set. The compounds and transition temperatures used in the development of the predictive models

(1) Sage, I. In *Critical Reports on Applied Chemistry (Vol. 22). Thermotropic Liquid Crystals*; Gray, G. W., Ed.; John Wiley & Sons: New York, 1987; Chapter 2.

(2) De Jeu, W. H.; Van Der Veen, J. *Mol. Cryst. Liq. Cryst.* **1977**, *40*, 1.

(3) Maier, W.; Saupe, A. *Naturforsch.* **1959**, *14a*, 882; **1960**, *15a*, 287.

(4) Knaack, L. E.; Rosenberg, H. M. *Mol. Cryst. Liq. Cryst.* **1972**, *17*, 171.

(5) Thiemann, T.; Volkmar, V. *Liq. Cryst.* **1997**, *22*, 519.

(6) Schröder, R.; Kränz, H.; Vill, V.; Meyer, B. *J. Chem. Soc., Perkin Trans.* **1996**, *2*, 1686.

(7) Kränz, H.; Vill, V.; Meyer, B. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1173.

(8) Wessel, M. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841.

(9) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28.

(10) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. *Anal. Chim. Acta* **1997**, *342*, 113.

(11) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S. *Anal. Chem.* **1994**, *66*, 1799.

(12) Sutter, J. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100.

(13) Engelhardt, H. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478.

(14) Liang, C.; Gallagher, D. A. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321.

(15) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879.

(16) Johnson, S. R.; Jurs, P. C. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H.; Testa, B.; Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997.

(17) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726.

(18) Cronin, M. T. D.; Dearden, J. C. *Quant. Struct.-Act. Relat.* **1995**, *14*, 518.

were taken from the literature.^{19–22} As discussed in each of the original sources, all transition temperatures were determined by optical microscopy using a Leitz Ortholux II POL BK microscope in conjunction with a Mettler FP 82 heating stage and FP 80 control unit. The transition temperatures were confirmed using a Mettler DTA TA 2000. The original papers do not report experimental errors for the clearing temperatures; however, an error estimate of a few degrees seems reasonable. In the development of the models here, the clearing temperature (in degrees Kelvin) was used as the dependent variable. The phase structure of the liquid crystal prior to the transition to an isotropic liquid was not considered during the model building process. The compounds and their experimental and calculated clearing temperatures are listed in Tables 1 and 2.

Each of the 318 calamatic compounds contained either a pyrimidine or a pyridine ring, one or two phenyl rings, and possibly a single cyclohexyl ring. Many of the cyclohexyl-containing compounds contained a linker group between the cyclohexyl ring and the adjoining phenyl ring. Due to the large difference in clearing temperatures between two- and three-ring core liquid crystals, the compounds were divided into two- and three-ring subsets for model construction. There were 209 two-ring compounds, with a clearing temperature range of 304–367 K and 109 compounds containing three rings, with a clearing temperature range of 364–492 K. Each of these structural subsets was divided into a training, cross-validation, and prediction set for the purpose of model building. The training set (tset) is used to train the models (i.e., adjust the weights and biases of a computational neural network). The cross-validation set (cvset) is used to determine the point at which a CNN has begun to learn tset-specific information and is no longer generalizing. The prediction set (pset) is used only for the validation of the final model. For the model developed for the entire dataset, the tset, cvset, and pset for both subsets were combined so that direct comparisons could be made among the three models.

Molecular Modeling. Each compound was sketched using HyperChem on a personal computer. The HyperChem model builder was used to place the structures in three-dimensional coordinates. As many of the descriptors that were calculated were dependent upon realistic geometries, the 318 structures were modeled more accurately using the semiempirical molecular orbital program MOPAC²³ with the PM3 Hamiltonian.²⁴ Due to the flexibility of the structures, many possible conformations were accessible. Structures were considered to be optimized when no conformation lower in energy could be found.

Descriptor Generation. A series of structural descriptors was calculated to numerically represent the structures. The descriptors calculated by ADAPT can be divided into three classes: geometric, electronic, and topological. Geometric descriptors, such as solvent ac-

cessible surface areas and moments of inertia, describe three-dimensional characteristics of the molecular structures. Topological descriptors are based on encoding molecular structures as graphs, consisting of nodes and edges. They include molecular connectivity indices, substructural descriptors, and path counts,^{25,26} which generally encode information regarding the size and degree of branching of the molecular structure. Electronic descriptors are calculated using an empirical atomic charge scheme.²⁷

A fourth class of descriptors which combine both electronic and geometric information was also calculated. These descriptors, known as charged partial surface area (CPSA) descriptors,²⁸ encode information regarding intermolecular interactions such as hydrogen bonding or polar interactions. Intermolecular interactions play a pivotal role in the liquid crystal phenomenon. As packing forces play a significant role in the liquid crystalline state, the standard hydrogen-bonding descriptors were extended to include weak hydrogen-bonding interactions. Specifically, a hydrogen bound to an electroneutral carbon was considered to be donatable, and phenyl rings were regarded as hydrogen-bond acceptors.²⁹ In an ordered, close-packed system (compared to an isotropic liquid or gaseous state), these intermolecular interactions likely play an important role in the lateral forces which must be overcome for the clearing transition.

Additional descriptors were suggested by work regarding chain ordering as an explanation of the odd-even effect.³⁰ The shorter and longer distance of the terminal atoms of the alkyl chains from the molecular axis defined by the ring-system core were calculated. These values were used as descriptors (e.g., DEVI), as were their sums and ratios. These descriptors are explained pictorially in Figure 1.

A number of other descriptors were designed with the specific application of liquid crystals in mind. For example, the fraction of the molecular mass that existed outside the rigid core of the liquid crystal was calculated for each structure. In addition, the mass of each atom in a alkyl/alkoxy/alkanoate chain was also scaled by its distance from the center of mass of the core structure (e.g., FLEX). In total, 218 descriptors were calculated for each structure.

Feature Reduction and Selection. The pool of descriptors was analyzed using two objective tests to reduce the number of features being considered during the model-building process. Descriptors containing 90% or more identical values were removed from the pool, as they contained little useful information. The remaining descriptors were then subjected to a pairwise correlation investigation. Pairs of descriptors correlating above 0.90 were identified, and one of the descriptors was eliminated from further consideration. For the three-ring subset of structures, this resulted in 48 descriptors remaining; 44 descriptors remained for the

(19) Kelly, S. M.; Fünfschilling, J. *J. Mater. Chem.* **1993**, 3, 953.

(20) Kelly, S. M.; Fünfschilling, J.; Villiger, A. *Liq. Cryst.* **1994**, 16, 813.

(21) Kelly, S. M.; Fünfschilling, J. *J. Mater. Chem.* **1994**, 4, 1673.

(22) Kelly, S. M.; Fünfschilling, J. *Liq. Cryst.* **1995**, 4, 519.

(23) Stewart, J. P. P. MOPAC 6.0, *Quantum Chemistry Program Exchange*; Indiana University: Bloomington, IN, Program 455.

(24) Stewart, J. P. P. *J. Comput.-Aided. Mol. Des.* **1990**, 69, 17.

(25) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, 24, 164.

(26) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley and Sons: New York, 1986.

(27) Dixon, S. L.; Jurs, P. C. *J. Comput. Chem.* **1992**, 13, 492.

(28) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, 62, 2323.

(29) Jeffrey, G. A. *An Introduction to Hydrogen Bonding*; Oxford University Press: New York, 1997.

(30) Marcelja, S. *J. Chem. Phys.* **1974**, 9, 3599.

Table 1. Observed and Calculated Clearing Temperatures for Structures Containing Two Rings in the Core

Obs.	n	R	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
 C_nH_{2n+1}					
1	7		331	330.4	326.0
2	7		353	347.4	349.5
3 ^c	7		304	316.7	317.3
4	7		324	323.0	319.6
5	7		304	311.5	322.6
6	7		328	316.1	321.0
7 ^a	7		319	315.5	322.3
8	8		329	332.0	326.7
9 ^a	8		348	347.0	349.6
10 ^c	8		326	315.2	314.2
11 ^b	8		323	318.7	322.6
12 ^c	8		307	316.0	325.9
13	8		325	323.4	325.8
14	8		317	317.9	324.5
15	9		338	337.0	331.3
16	9		352	347.2	355.0
17 ^{b,c}	9		330	318.4	318.7
18 ^c	9		334	325.6	325.5
19 ^c	9		317	321.9	328.8
20 ^b	9		333	328.7	325.1

Table 1. (Continued)

Obs.	n	R	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
21	9		329	327.4	327.3
22	9		348	347.0	347.3
23	9		346	347.8	346.0
24	9		327	337.0	330.3
25	9		346	342.2	341.2
26	9		334	337.1	345.1
27	9		345	343.7	341.2
28 ^b	9		340	339.4	349.0
 C_nH_{2n+1}					
29	7		350	353.7	345.1
30	7		360	356.7	358.2
31	7		329	329.3	332.7
32	7		346	342.9	337.8
33	7		329	328.3	340.4
34 ^a	7		343	337.1	338.7
35 ^b	7		340	336.8	339.7
36	8		353	353.0	351.5
37	8		356	356.5	362.4
38	8		332	334.2	335.3
39 ^b	8		349	335.7	342.1
40	8		332	334.1	344.8

Table 1. (Continued)

Obs.	n	R	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
41 ^b	8		346	342.1	344.0
42	8		343	342.5	342.0
43	9		359	358.6	353.9
44	9		359	356.8	361.2
45 ^c	9		338	334.6	335.9
46 ^a	9		357	342.3	344.3
47	9		338	342.1	343.3
48	9		353	348.8	342.8
49	9		349	350.3	346.6
50	9		358	357.5	362.5
51	9		361	361.1	359.2
52	9		339	336.7	332.3
53	9		359	357.6	355.8
54	9		344	337.2	347.8
55 ^b	9		357	357.9	355.2
56	9		351	351.1	350.5
Obs.	n	m	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
57 ^c	7	1	360	356.9	345.9
58 ^c	7	2	342	340.3	341.7
59	7	3	355	350.1	345.3
60	7	4	345	348.6	344.7
61	7	6	349	346.9	347.7
62	7	7	353	348.0	349.7
63	7	8	347	349.6	348.7
64	7	9	353	352.8	350.2
65 ^a	8	1	353	357.4	353.8
66 ^c	8	2	337	341.3	344.1

Table 1. (Continued)

Obs.	n	m	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
67	8	3	349	348.0	347.5
68	8	4	339	346.4	345.9
69	8	6	343	346.2	348.3
70	8	7	349	347.7	350.4
71	8	8	347	350.7	349.9
72	8	9	351	354.0	351.3
73 ^b	9	1	362	358.5	362.5
74 ^c	9	2	343	345.1	348.5
75 ^a	9	3	354	348.9	351.9
76	9	4	343	347.3	348.5
77	9	6	349	347.6	350.0
78	9	7	354	349.8	352.1
79	9	8	351	352.3	351.0
80	9	9	355	355.9	352.6
81	7	1	366	360.4	356.9
82	7	2	348	346.0	350.8
83	7	3	360	355.4	353.7
84	7	4	352	356.3	352.1
85	7	6	356	357.9	353.2
86	7	7	360	358.8	355.6
87	7	8	358	359.8	354.1
88 ^b	7	9	360	361.0	356.5
89	8	1	362	359.9	357.9
90	8	2	345	359.9	357.7
91	8	3	356	354.9	357.5
92	8	4	348	355.6	354.7
93 ^a	8	6	353	357.4	354.6
94 ^a	8	7	358	358.3	357.5
95 ^b	8	8	356	359.4	355.4
96	8	9	359	360.9	357.7
97	9	1	367	360.5	364.2
98 ^a	9	2	350	348.7	357.9
99	9	3	360	355.3	362.4
100	9	4	353	355.8	356.6
101	9	6	356	357.4	357.2
102	9	7	361	358.4	360.4
103	9	8	360	359.9	357.7
104	9	9	363	361.3	361.6
Obs.	n	R	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
105	6		358	358.4	349.9
106 ^b	6		353	349.8	345.5
107 ^c	6		328	329.2	339.4
108	6		353	348.5	346.9

Table 1. (Continued)

Obs.	n	R	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
109	6		328	342.9	339.2
110	6		358	352.4	345.6
111	6		348	352.6	349.7
112	7		360	358.9	352.6
113	7		355	351.6	348.4
114 ^c	7		328	326.8	346.1
115	7		355	346.2	349.2
116	7		328	339.2	341.9
117	7		360	351.8	348.7
118	7		351	352.2	350.7
119	8		357	359.0	355.0
120 ^a	8		357	352.6	350.5
121 ^{b,c}	8		316	326.1	351.9
122	8		355	343.0	349.2
123	8		330	337.8	350.6
124 ^a	8		358	348.9	349.1
125	8		349	352.3	352.5
Obs.	n	m	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
126	6	2	336	336.1	339.6
127	6	3	349	347.8	341.8
128 ^b	6	4	344	348.9	344.9
129	6	6	351	350.0	349.4
130	6	7	355	352.1	350.8
131	6	8	355	354.9	353.6

Table 1. (Continued)

Obs.	n	m	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
132	6	9	357	357.4	354.8
133	7	2	339	337.9	339.9
134 ^a	7	3	351	346.6	346.3
135	7	4	346	350.3	348.3
136	7	6	353	353.7	351.3
137	7	7	359	354.7	353.2
138 ^a	7	8	358	355.6	354.2
139 ^b	7	9	360	356.0	355.2
140	8	2	338	337.7	344.4
141	8	3	350	347.5	349.6
142	8	4	346	349.9	350.6
143 ^b	8	6	352	354.5	355.8
144	8	7	357	356.6	356.6
145	8	8	358	358.5	356.4
146	8	9	359	359.6	360.9
147	9	2	339	339.8	346.4
148	9	3	350	348.3	345.9
149	9	4	346	351.3	353.1
150 ^a	9	6	353	355.9	357.4
151	9	7	357	357.7	359.5
152	9	8	359	359.4	361.4
153 ^a	9	9	360	360.5	362.5
154	6	3	342	340.5	340.6
155 ^b	6	4	337	346.5	345.5
156	6	5	352	350.0	345.6
157	6	7	354	353.5	350.1
158	6	8	353	354.7	352.3
159	6	9	355	355.5	353.3
160	6	10	354	356.1	354.5
161	7	3	345	340.5	341.1
162	7	4	340	346.3	346.7
163	7	5	354	349.9	347.4
164	7	7	357	353.8	351.9
165	7	8	355	355.1	353.8
166	7	9	358	356.1	354.9
167	7	10	357	356.8	355.7
168 ^b	8	3	342	340.8	342.7
169	8	4	337	346.0	348.6
170	8	5	353	349.6	348.9
171	8	7	356	353.9	353.5
172 ^a	8	8	354	355.2	354.9
173	8	9	357	356.3	356.3
174	8	10	355	357.0	357.0
175	9	3	343	341.0	344.8
176 ^a	9	4	338	345.8	349.8
177	9	5	354	349.4	351.2
178 ^a	9	7	357	354.0	355.2
179	9	8	355	355.5	356.5
180	9	9	358	356.8	357.5
181 ^a	9	10	356	357.7	357.9
182	6	5	335	335.8	344.3
183	6	6	350	350.3	342.4
184	6	7	350	356.2	349.7
185	6	9	357	358.0	354.1

Table 1. (Continued)

Obs.	n	m	Obs CT (K)	Calc CT ^d (K)	Calc CT ^e (K)
186	6	10	360	358.1	355.2
187 ^a	6	11	359	358.9	357.7
188 ^b	6	12	360	359.8	359.2
189	7	5	339	334.7	346.1
190	7	6	351	348.4	345.0
191 ^b	7	7	351	355.6	351.7
192	7	9	358	359.3	356.3
193	7	10	361	359.6	357.4
194	7	11	362	360.3	359.6
195	7	12	362	360.8	361.1
196	8	5	334	334.5	348.1
197 ^a	8	6	349	347.6	347.8
198	8	7	351	355.1	353.7
199	8	9	358	360.1	358.4
200	8	10	361	360.6	359.8
201	8	11	360	361.2	361.3
202	8	12	362	361.6	363.1
203	9	5	336	335.1	350.4
204 ^a	9	6	350	347.2	350.6
205 ^b	9	7	351	355.2	356.0
206	9	9	358	360.7	360.4
207	9	10	360	361.5	362.2
208 ^b	9	11	360	362.2	363.0
209	9	12	362	362.7	364.2

^a Member of the cross-validation set (cvset). ^b Member of the prediction set (pset). ^c Monotropic transition or no liquid crystalline phase identified (melting point used). ^d Calculated clearing temperatures from the core-specific model. ^e Calculated clearing temperatures from the combined-sets model.

two-ring subset of structures. For the combined model, 66 descriptors remained after the objective feature reduction. Topliss and Edwards³¹ have suggested that the ratio of the number of descriptors to the number of observations be kept below 0.60 to lessen the probability of chance correlations in a QSPR model. Each of these reduced pools meet this criterion for the number of descriptors versus the number of available training observations.

Following this objective feature reduction, the reduced pools of descriptors were screened using subjective feature selection methods to identify subsets of descriptors which related well to the liquid crystal clearing temperatures. A genetic algorithm routine was used to evaluate subsets of descriptors using a PRESS (predictive residual error sum of squares) fitness from a CNN training as the measure of subset quality. The genetic algorithm finds a small, information rich subset of descriptors using a directed search of the feature space. The starting weights and biases for each CNN training were selected using a generalized simulated annealing optimization method.³² Subsets with fewer descriptors but equivalent root-mean-square (rms) errors were favored in the model selection process. The best subsets identified using this approach were retained for further analysis regarding the architecture of CNN for the fully trained final model.

Computational Neural Networks. For the purposes of QSPRs, a CNN can be considered as a nonlinear

regression method. Neural networks relate a series of input variables to a targeted output through a process known as training. There are several different methods used in training a fully connected, three-layer, feed-forward neural network. The CNN used for the work in this paper utilizes a quasi-Newton method (BFGS)³³⁻³⁷ to optimize the weights and biases.³⁸

To avoid overtraining, a cross-validation set (cvset) was used periodically throughout the training process to evaluate the networks' ability to predict using the current weights and biases. During the early stage of training, the cvset rms error will decline along with the rms error of the training set (tset). Eventually the rms error of the cvset will begin to rise. At this point, the network is no longer learning general information; rather, it has begun to memorize information specific to the training data. The set of weights and biases corresponding to the cvset rms error minimum were retained and used for later prediction.

The determination of the best CNN architecture was done empirically. A small number of hidden neurons was used at the beginning. Additional hidden neurons were added stepwise until no appreciable improvement in the tset and cvset rms errors was found. However, the number of adjustable parameters in the final network architecture should never exceed a number greater than half the number of training observations

(33) Broyden, C. G. *J. Inst. Maths. Appl.* **1970**, *6*, 76.

(34) Fletcher, R. *Comput. J.* **1970**, *13*, 317.

(35) Goldfarb, D. *Math. Comput.* **1970**, *24*, 23.

(36) Shanno, D. F. *Math. Comput.* **1970**, *24*, 647.

(37) Fletcher, R. *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization*; Wiley: New York, 1980.

(38) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. *Environ. Toxicol. Chem.* **1994**, *13*, 841.

(31) Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238.

(32) Sutter, J. M.; Jurs P. C. Selection of Molecular Structure Descriptors for Quantitative Structure-Activity Relationships. In *Adaptation of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science Publishers B. V.: Amsterdam, 1995.

Table 2. Observed and Calculated Clearing Temperatures for Structures with Three Rings in the Core

Obs.	N _{Pos}	R	Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)
210	W		456	451.2	461.8
211 ^a	W		460	461.0	458.7
212	W		440	451.8	441.8
213	W		459	450.8	449.0
214	W		445	451.3	451.5
215	W		457	450.6	450.8
216	W		451	447.8	453.3
217	X		478	469.4	476.4
218	X		484	479.8	480.1
219	X		458	470.5	461.5
220	X		483	469.7	470.5
221	X		465	471.0	473.0
222	X		478	472.3	473.7
223	X		473	472.3	473.6
224	Y		468	463.0	461.4
225	Y		468	471.9	473.2
226	Y		457	463.7	460.8
227 ^a	Y		469	463.1	470.8
228	Y		460	463.6	471.5
229	Y		467	463.4	470.8

Table 2. (Continued)

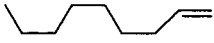
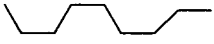
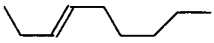
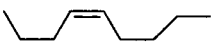

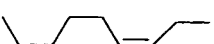
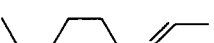
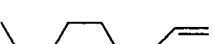
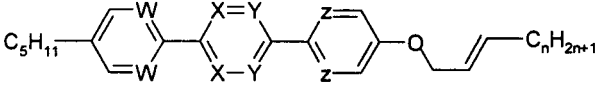
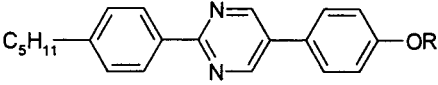
Obs.	N _{Pos}	R	Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)
230	Y		463	463.7	464.1
231	Z		460	456.5	454.6
232 ^a	Z		458	459.4	459.3
233 ^b	Z		439	456.1	439.7
234	Z		461	455.9	453.5
235	Z		446	456.1	451.3
236	Z		461	456.0	454.2
237	Z		456	455.4	453.6
					
238	W	3	469	467.6	461.8
239	W	4	461	464.5	460.5
240	W	6	453	457.4	455.6
241	W	7	451	453.2	456.0
242	W	8	445	448.4	451.2
243 ^b	W	9	442	442.9	452.1
244	X	3	492	481.8	483.0
245	X	4	488	480.9	480.0
246	X	6	479	477.5	475.1
247	X	7	477	475.2	474.8
248	X	8	474	472.2	469.1
249	X	9	471	468.9	468.5
250	Y	3	476	472.3	479.1
251	Y	4	469	471.1	465.5
252	Y	6	465	467.2	461.9
253	Y	7	463	464.1	465.2
254	Y	8	460	459.8	455.1
255	Y	9	458	454.3	456.3
256	Z	3	468	463.6	463.5
257 ^b	Z	4	460	461.9	458.3
258	Z	6	452	455.5	455.9
259	Z	7	450	450.9	454.7
260	Z	8	446	445.5	452.9
261	Z	9	445	440.4	449.4
					
262 ^b		Hexyl	484	473.8	482.2
263		(Z)-Hex-2-enyl	401	416.4	403.9
264		(E)-Hex-3-enyl	443	474.5	476.1
265		(Z)-Hex-3-enyl	470	475.3	470.6
266		(E)-Hex-4-enyl	490	474.4	476.9
267 ^a		(Z)-Hex-4-enyl	455	475.6	464.1

Table 2. (Continued)

Obs.	N_{Pos}	R		Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)	
268 ^a		Hex-5-enyl		479	472.8	471.8	
Obs.	n	m	R	Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)	
269	3	7	—	442	441.2	438.2	
270 ^b	5	7	—	435	440.6	435.8	
271	7	7	—	431	433.5	435.3	
272	3	7	C ₂ H ₄	416	419.7	416.6	
273 ^a	5	7	C ₂ H ₄	413	417.7	408.8	
274 ^b	7	7	C ₂ H ₄	410	409.3	394.3	
275	3	7	CH ₂ O	424	421.7	426.8	
276	5	7	CH ₂ O	425	424.0	422.5	
277	7	7	CH ₂ O	417	420.4	415.4	
278	3	7	CO ₂	454	436.3	450.2	
279	5	7	CO ₂	453	440.8	452.7	
280	7	7	CO ₂	443	441.4	444.9	
281	3	10	—	424	420.1	422.7	
282 ^b	5	10	—	425	420.7	420.8	
283	7	10	—	422	420.7	421.2	
284	3	10	C ₂ H ₄	404	401.0	408.5	
285	5	10	C ₂ H ₄	405	401.6	405.4	
286	7	10	C ₂ H ₄	403	402.0	400.3	
287	3	10	CH ₂ O	412	407.8	408.5	
288	5	10	CH ₂ O	415	413.1	412.5	
289	7	10	CH ₂ O	409	408.7	409.6	
290	3	10	CO ₂	434	430.9	427.4	
291	5	10	CO ₂	434	436.1	434.2	
292	7	10	CO ₂	431	433.2	432.5	
293	1	10	CO ₂	397	421.0	396.9	
294	2	10	CO ₂	415	431.4	428.6	
295	4	10	CO ₂	433	433.1	435.7	
296	9	10	CO ₂	426	428.8	427.8	
297	10	10	CO ₂	423	426.8	419.1	
Obs.	R ₁		Z	R ₂	Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)
298 ^b			CO ₂	C ₁₀ H ₂₁	435	430.7	428.0
299			CO ₂	C ₁₀ H ₂₁	423	432.2	426.5
300			CO ₂	C ₁₀ H ₂₁	443	436.4	432.2
301	C ₄ H ₉ O		-	C ₁₀ H ₂₁	416	415.0	418.0
302	C ₃ H ₇		OCH ₂	C ₁₀ H ₂₁	364	361.0	365.4
303	C ₅ H ₁₁		-	OC ₉ H ₁₉	449	444.9	446.2
304	C ₄ H ₉ O		-	OC ₉ H ₁₉	440	448.9	455.2
305	C ₂ H ₅ O		CH ₂ O	C ₁₀ H ₂₁	403	402.7	398.8
306	C ₃ H ₇		CH ₂ O	OC ₉ H ₁₉	431	438.5	438.0

Table 2. (Continued)

Obs.	R ₁	Z	R ₂	Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)
307	C ₂ H ₅ O	CH ₂ O	OC ₉ H ₁₉	427	427.0	429.7
308	C ₃ H ₇ CO ₂	-	C ₁₀ H ₂₁	431	434.2	429.6
309	C ₃ H ₇	O ₂ C	C ₁₀ H ₂₁	415	409.4	416.5
310	C ₅ H ₁₁	-	O ₂ CC ₈ H ₁₇	451	450.8	451.4
311	C ₃ H ₇	CH ₂ O	O ₂ CC ₈ H ₁₇	442	438.0	438.7
Obs.	R			Obs CT (K)	Calc CT ^c (K)	Calc CT ^d (K)
312 ^b				454	448.8	447.6
313 ^a				449	450.9	451.2
314 ^b				434	448.9	435.3
315				457	448.4	454.6
316				441	448.9	439.1
317				458	448.4	451.2
318 ^a				449	446.3	455.2

^a Member of the cross-validation set (cvset). ^b Member of the prediction set (pset). ^c Calculated clearing temperatures from the core-specific model. ^d Calculated clearing temperatures from the combined-sets model.

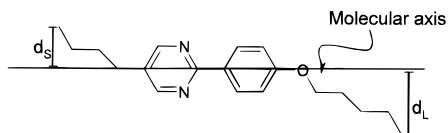


Figure 1. Schematic illustration of the calculation of the largest and smallest deviation from the molecular axis. The molecular axis is defined as the line connecting the first atoms on either side of the rigid core structure.

used in the training process. In other words, the ratio of the number of tset members to the number of adjustable parameters in the CNN³⁹ should always be greater than 2.0. The final, fully trained CNN was then validated using the external prediction set as an estimate of the utility of the model for predictions.

Descriptor generation, model development, and all statistical analysis was performed using the Automated Data Analysis and Pattern Recognition Toolkit (ADAPT) developed in our laboratory^{40,41} and running on a DEC Alphastation 500.

Results and Discussion

Clearing Temperature of Two-Ring Core Structures. The available data for the development of a QSPR model to predict the clearing temperatures of compounds containing two rings in the core were divided into a training set of 165 compounds and cross-validation and prediction sets of 22 compounds each. As described above, the training set was used to adjust the weights and biases on the CNN, while the cvset was used to determine an adequate stopping point for the training process. The pset was used only to validate the fully trained model.

The seven descriptors selected using the GA-CNN feature selection tool with a leave-20%-out PRESS cost function are shown in Table 3. The pairwise correlation coefficient among these descriptors has an average value of 0.28 and a range of 0.04–0.68. Three of the descriptors chosen for this model were topological. The molecular distance edge descriptor⁴² encodes information regarding the length of the alkyl/

(39) Livingstone, D. J.; Manallack, D. T. *J. Med. Chem.* **1993**, *36*, 1295.

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

(41) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, D. C., 1979; pp 103–129.

(42) Liu, S.; Chenzhong, C.; Li, Z. *J. Chem. Inf. Sci.* **1998**, *38*, 387.

Table 3. Features Chosen Using GA-CNN for the Prediction of the Clearing Temperature of Structures with Two Rings in the Rigid Core

descriptor	label
atomic charge weighted partial positive surface area	PPSA-3
weighted mass of flexible chains	FLEX
valence corrected path-6 clusters	V6PC
valence corrected path-7 chains	V7CH
molecular distance edge-14	MDE-14
average charge on H-bond acceptor atoms	ACAA
$\max(q_H - q_A)$	MCHG

alkoxy chains:

$$d_{14} = \prod_{N_{14}} (L_{i1,j4})^{1/(2N_{14})}$$

where $L_{i1,j4}$ is the path length between the i th carbon bound to three hydrogens (*primary*) and the j th carbon bound to no hydrogens (*quaternary*), and N_{14} is the total number of paths between primary and quaternary carbons. Additionally, two valence-corrected graph theoretic descriptors were chosen for use in the model. These descriptors are encoding additional information regarding the length of the side chains, as well as the location of any double bonds and the nature of alkoxy or alkanooate linkages to the core. The weighted mass of the flexible chains consists of the sum of the atomic weights of the atoms in the side chains, where each atom is weighted by its distance from the center-of-mass of the rigid core of the molecule. The number of descriptors present in the model which explicitly encode information regarding the nature of the alky/alkoxy/alkanoate chains indicates the substantial role that these moieties play in the clearing temperature. This is not surprising, as the rigid core of the structures varies only slightly over the two-ring portion of the data set.

Additional descriptors encode the ability of the molecules to undergo hydrogen bonding. As mentioned in the Experimental Section, the hydrogen-bonding descriptors were extended to include very weak H-bond interactions which are normally ignored in ADAPT hydrogen-bonding studies. One hydrogen-bonding descriptor (ACAA) contains information regarding the ability to accept a hydrogen bond. MCHG is the maximum charge difference between a donatable hydrogen and the atom to which it is formally bound (the hydrogen donor, A), a measure of the hydrogen-bond strength.

A plot of calculated versus observed clearing temperature for this model is shown in Figure 2. The pairwise correlation coefficient between calculated and observed clearing temperature for all 209 compounds is 0.928, and for the 22 compounds in the prediction set it is 0.911. The clearing temperature range shown in Figure 2 is from 300 to 370 K. Generally speaking, the model appears best in the region which contains the bulk of the available data. As the data become more sparse, the observations tend to deviate more from the 1:1 correlation line. This trend is not surprising, as empirical methods tend to work best when a high observation density is available in the training data. Nonetheless, the model does encode the transition temperatures quite well. The similar training set and prediction set rms errors indicate that the model is robust and is capable

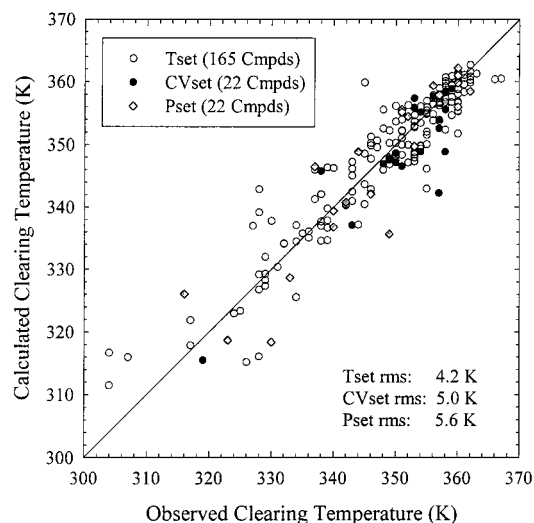


Figure 2. Calculated clearing temperatures versus the observed clearing temperatures for a structures containing two rings. Compounds that have a calculated value exactly equal to the observed value would lie on the 1:1 correlation line shown.

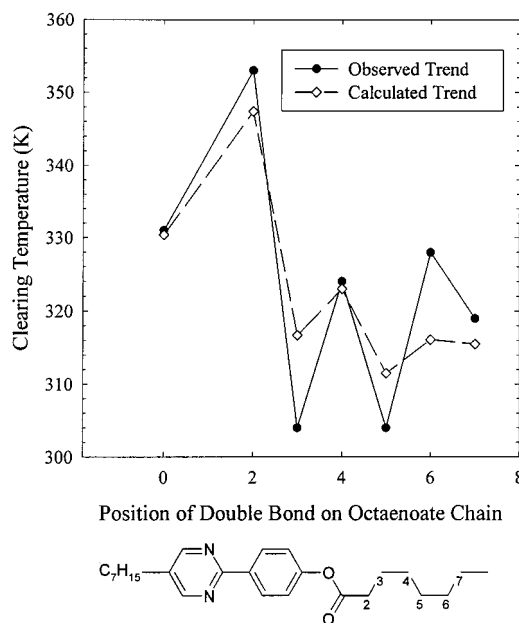


Figure 3. Trend in a homologous series of compounds as the number of carbons in the alkoxy chain is increased. While the calculated clearing temperatures correspond quite well to those observed, the odd-even effect is not well-encoded in this model.

of making accurate predictions for compounds that are structurally similar to the training compounds.

An interesting measure of the quality of a model for a data set such as this is the degree to which it can capture the trends in homologous series of compounds. Figure 3 shows the clearing temperature trend over a series of structures as the alkoxy chain length is increased by the addition of carbons to the chain end. The observed trend shows the well-known odd-even effect, in which an oscillation in the clearing temperature is seen as each additional carbon is added to the chain. Although there is no consensus in the literature over the cause of this trend, it has been suggested that it is correlated with the molecular polarizability anisotropy, $\Delta\alpha$ ($\alpha_{||} - \alpha_{\perp}$). Although the actual clearing

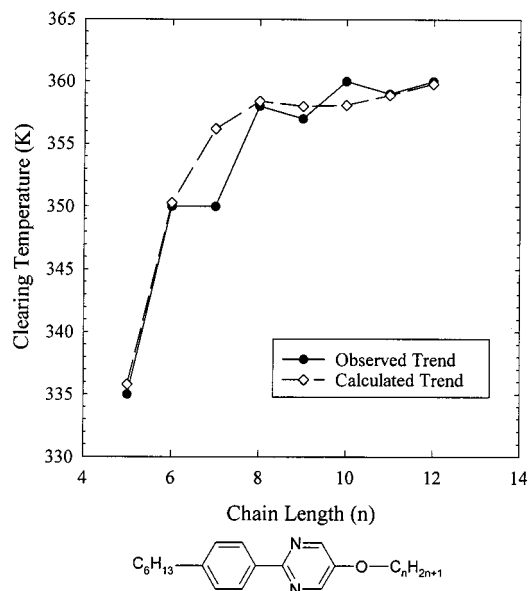


Figure 4. Trend in a homologous series of compounds as the position and orientation of the double bond in an octanoate chain is altered. In this example, the oscillations in the clearing temperatures are well-captured by the fully trained CNN.

Table 4. Features Selected Using GA-CNN for the Prediction of the Clearing Temperatures of 3-Ring Core Liquid Crystals

descriptor	label
relative negative charge	RNCG
average charge on H-bond acceptor atoms	ACAA
distance between most positive and negative atomic σ charges	CSEP
number of path-6 clusters	S6PC
total number of paths/number of atoms	ALLP
$(\sum_{\text{don-H}} q_{\text{H}} - q_{\text{A}} \times SA_{\text{H}} \times q_{\text{H}}) / SA_{\text{tot}}$	WCDH

temperatures are matched quite well by the calculated values, the odd–even effect is not well-captured by the model in this case. Figure 4 shows another trend, as the position of a carbon–carbon double bond is moved down the octanoate chain. In this case, the oscillatory trend seen in the observed data is represented quite well in the calculated trend, although the calculated oscillations are somewhat smaller than those observed.

Clearing Temperature of Three-Ring Core Structures. A separate model was developed for compounds with three rings in their cores. These compounds contained more diversity in the ring structure compared to the two-ring structures. Again, the data set was divided in to a training set, cross-validation set, and external prediction set. A six-descriptor model was identified using the GA-CNN feature selection routine, and it was retrained using a 6–5–1 CNN. The six descriptors chosen are shown in Table 4. Pairwise correlations among these descriptors range from 0.03 to 0.85, with an average value of 0.26. The calculated versus observed clearing temperature plot, with a range of 340–500 K is shown in Figure 5. The pairwise correlation coefficient between calculated and observed clearing temperature for all 109 compounds is 0.949, and for the 10 compounds in the prediction set it is 0.903.

Two of the six descriptors present in this model are topological, and they encode information regarding the length of the side chains. The distance between the most positive and negative σ charges largely encodes the

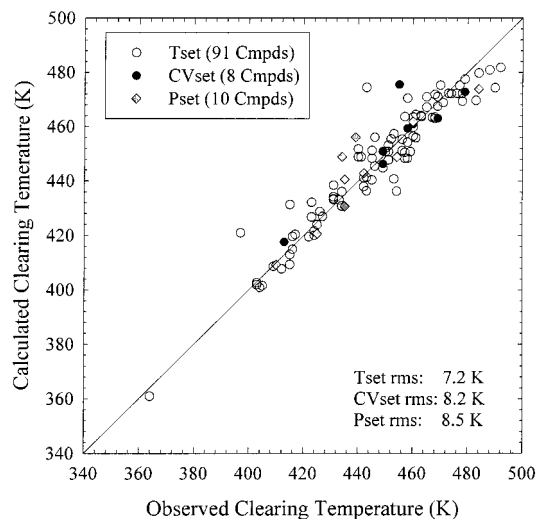


Figure 5. Calculated clearing temperatures versus the observed clearing temperatures for a structures containing three rings.

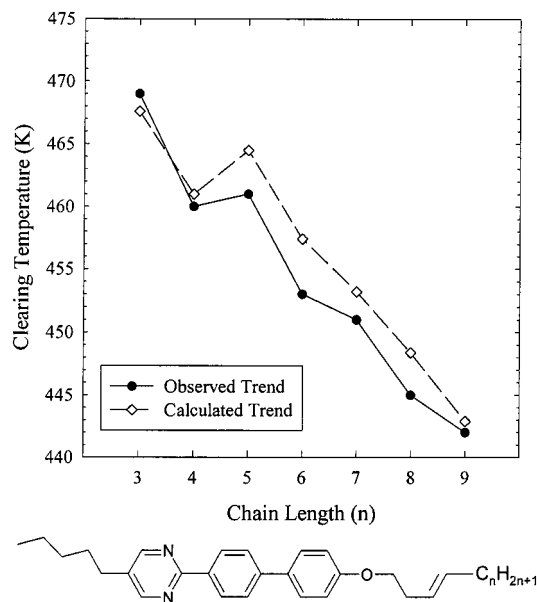


Figure 6. Homologous series of compounds in which the number of carbons in the alkenoxy chain is increased.

position and orientation of the pyrimidine ring in the central core structure. The relative negative charge, average charge on H-bond acceptor atoms, and the weighted charge on donatable hydrogens are somewhat more difficult to interpret, but seem to relate to the presence of an alkoxy linkage and the position of the double bond on the side chains.

Several plots demonstrating the ability of the model to encode information regarding trends in homologous series are presented in Figures 6 and 7. As with the two-ring core model, many of the trends in the experimental clearing temperatures are seen in the calculated clearing temperatures. These figures show a few typical examples of homologous trends in the three-ring core structures.

Combined Sets Model. To facilitate comparison between models, the tset, cvset, and pset for development of this model were each composed of a combination of the respective sets from the development of the subset models.

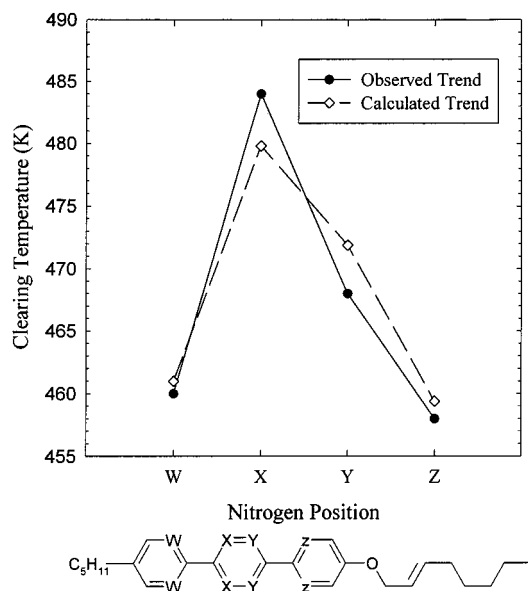


Figure 7. Trend in the observed and calculated clearing temperatures as the location of the pyrimidine ring is altered.

Table 5. Features Selected Using GA-CNN for the Prediction of the Clearing Temperatures of All Liquid Crystals Listed in Table 1

descriptor	label
relative positive charge	RPCG
fractional atomic charge weighted partial negative surface area	FNSA-3
atomic charge weighted partial positive surface area	PPSA-3
average charge on H-bond acceptor atoms	ACAA
distance between most positive and negative atomic σ charges	CSEP
number of path-3 clusters	S3C
valence corrected path-5 cluster	V5C
molecular distance edge-24	MDE-24
distance of the farthest terminal alkyl atom from the molecular axis defined by the ring structure	DEVI
$(\sum_{\text{don-H}} q_{\text{H}} - q_{\text{A}} \times SA_{\text{H}} \times q_{\text{H}}) / SA_{\text{tot}}$	WCDH

The best model identified by the GA-CNN feature selection routine contained 10 descriptors, which are listed in Table 5. This model was identified using a leave-25%-out PRESS validation method as the cost function in the genetic algorithm feature selection routine. Pairwise correlations between the 10 descriptors range from a low of 0.01 to a high of 0.88, with an average value of 0.31. The final network architecture was 10-5-1. Figure 8 shows the calculated clearing temperatures versus the experimentally measured clearing temperatures. The pairwise correlation coefficient between calculated and observed clearing temperature for all 318 compounds is 0.992, and for the 30 compounds in the prediction set it is 0.983. The pset rms error is lower than that for the three-ring model and slightly larger than for the two-ring model. Calculating the error for only the three-ring structures yields a marked improvement over the rms error derived from the three-ring specific model. This would indicate that there is additional structural information obtained by using all of the compounds in the development of the model. Indeed, the rms error for the two-ring structures in the pset also shows an improvement over the two-ring only model.

An exception to this improvement would be the compounds that are monotropic liquid crystals. Mono-

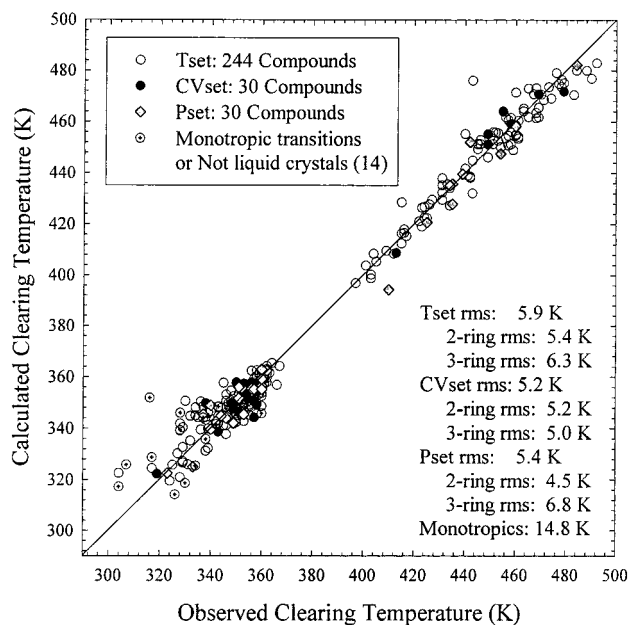


Figure 8. Calculated clearing temperatures versus the observed clearing temperatures for all structures used in this study.

otropic liquid crystals enter the liquid crystalline phase only when the temperature changes in one direction. Typically, monotropic materials have a liquid crystal transition temperature below the melting temperature, so the liquid crystalline phase can only be entered through supercooling. During the training of this combined model, it was noticed that the monotropic materials had an rms error of approximately 11 K, well above the error for the other compounds. For this reason, they were withheld during the training phase, and their clearing temperatures were predicted using the fully trained model. They were present, however, during the feature selection process.

Glancing at the descriptors present in this model, it is obvious that a substantial amount of overlap exists between this model and the two core-specific models described previously. For example, the average charge on hydrogen-bond acceptors (ACAA) is present in all three of the clearing temperature models. The distance between the most positive and negative atomic σ charges (CSEP) is also present in the three-ring core model, as is the weighted charge of donatable hydrogens (WCDH). PPSA-3 is present in both the combined model and the two-ring core model. Also present in each of the models are a number of topological path-length descriptors. The large degree of overlap between the three models may help explain the improvement in the pset rms error.

Figures 9 and 10 demonstrate the ability of this combined model to encode the trends in clearing temperatures in homologous trends. As with the core-specific models many, but not all, of the trends are well-captured by the empirical model. The ability to predict the trend among a series of homologues may prove beneficial in some application design scenarios.

Conclusion

A series of QSPR models has been developed using the genetic algorithm and computational neural net-

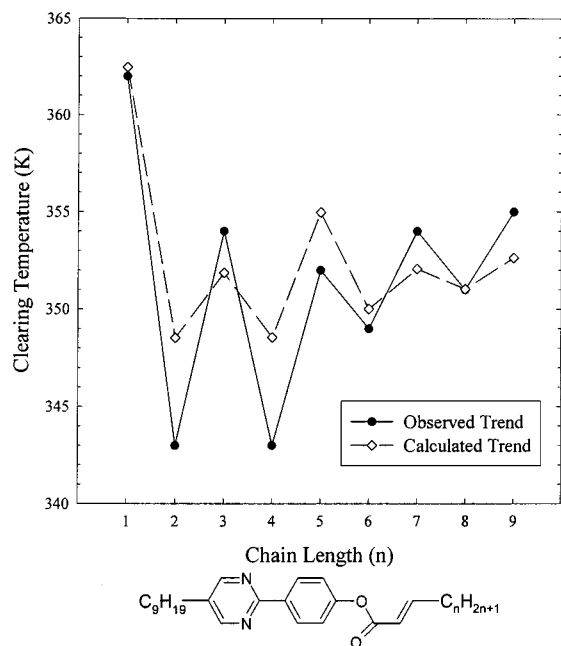


Figure 9. Trend in the observed and calculated clearing temperatures as the number of carbons is increased in the alkenoate chain.

works for the prediction of liquid crystal clearing temperatures. While the models do not predict the clearing temperatures to within experimental error, many of the trends in homologous series are captured by the models. There are several areas in which improvement could yield significant improvements in the prediction errors for this type of application. The first, and likely most influential, is the single conformer selected from which to generate descriptors. The structures used in this work were modeled to a low-energy conformation in vacuo. Of course, with structures that are as inherently flexible as those used here, a single low-energy conformer is only an approximation of the dynamic range of conformers present in the liquid crystalline phase. Additionally, the encoding of packing forces in close-packed systems, such as molecular crystals, has long been a stumbling block in the prediction

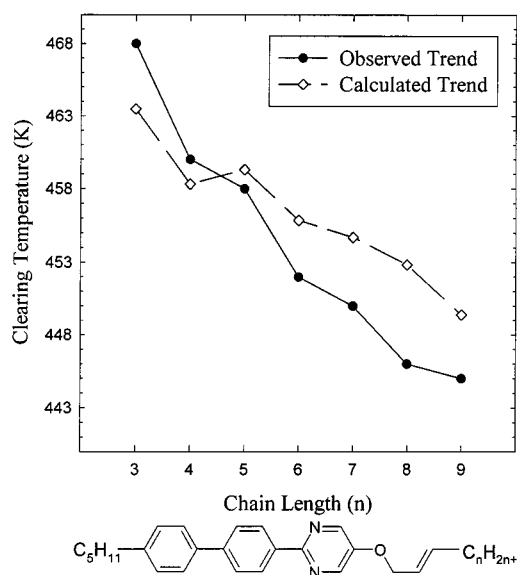


Figure 10. Trend in the observed and calculated clearing temperatures as the number of carbons is increased in the alkenoate chain.

of properties such as melting points.⁴³ It is likely that improvements in this area would enhance the prediction of liquid crystalline properties as well.

The QSPR models developed here included numerical descriptors regarding structural topology and the nature of any intermolecular interactions required for thermostability. These features, coupled with computational neural networks, were capable of predicting the clearing temperatures with an rms error of 5.4 K for a series of compounds not used to select descriptors or to train the neural networks. Future work should investigate using this QSPR methodology to investigate relationships among materials with a more diverse range of structural features.

CM980674X

(43) Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913.