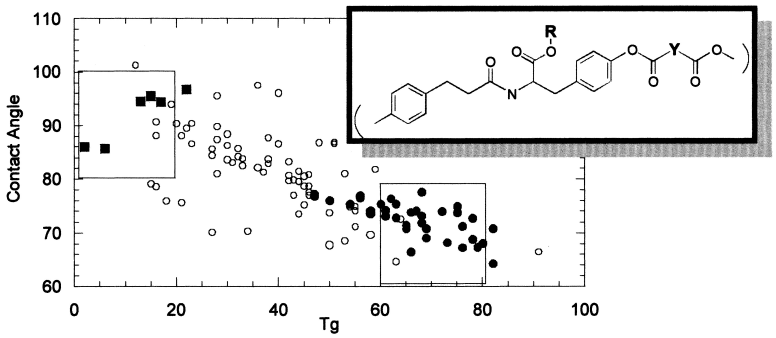
JOURNAL OF combinatorial CHEMISTRY

Article

Designing Diverse and Focused Combinatorial Libraries of Synthetic Polymers

Charles H. Reynolds

J. Comb. Chem., **1999**, 1 (4), 297-306• DOI: 10.1021/cc9900044 • Publication Date (Web): 05 June 1999



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 2 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Combinatorial Chemistry is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

Designing Diverse and Focused Combinatorial Libraries of Synthetic Polymers

Charles H. Reynolds*

Rohm and Haas Company, 727 Norristown Road, Spring House, Pennsylvania 19477-0904

Received February 12, 1999

Molecular topology and genetic algorithm optimized quantitative structure−property relationships (QSPR) have been used to design diverse and focused libraries of synthetic biodegradable polymers. A diverse subset (17 polymers) of a 112-member virtual polymer library was selected based on the molecular topology of the repeat unit using a stochastic diversity method (SimSearch-SCA). These 17 polymers were shown to be highly representative of the two-dimensional property space for the full library where the properties of interest are glass transition temperature (Tg) and hydrophobicity as measured by the air−water contact angle (CA). The 17 polymers in the diverse library were used to derive QSPR equations for Tg and CA by using a genetic algorithm to select molecular topology descriptors for linear regression. High quality models were derived for both Tg and CA. These QSPR models were tested by comparing the computed and experimental Tg and CA values for the 95 polymers that were *not* included in the training set. Representative models give r^2 values of 0.89 and 0.92 for Tg and CA, respectively. The QSPR models were further tested by using them to build focused libraries with specific values of Tg and CA. The focused libraries were very successful in identifying polymers that fall within specified ranges of Tg and CA. This work illustrates that the same concepts of molecular similarity and diversity that have been exploited so effectively in the pursuit of small biologically active molecules can also be employed in the design of synthetic polymers, particularly in the context of parallel synthesis.

Introduction

The automated synthesis of libraries of small molecules using combinations of synthetic building blocks has become an important strategy for dramatically increasing the number of compounds available for high throughput screening, particularly in pharmaceuticals and agrochemicals.[1,2] Despite recent gains in the rapid synthesis and testing of new compounds, it has become increasingly apparent that it is still impossible to simply make everything. This realization has led to greater use of rational approaches[3−10] to design combinatorial libraries for synthesis. The principal consideration in designing compound libraries for synthesis is whether the objective is a diverse library for screening or a focused library for rapid synthesis of analogues. Libraries may also be designed to fulfill other criteria such as ease of synthesis or reagent efficiency.

In comparison to pharmaceuticals[11] and agrochemicals,[12] combinatorial approaches have only recently been applied in the design of new materials.[13,14] Examples of material science problems that have been tackled using parallel synthesis include catalysts,[15−17] phosphors,[18] superconductors,[19] electronic materials,[20−22] and polymers.[23,24] In many ways polymers would seem to be particularly amenable to parallel synthesis given the wide variety of monomers that can be readily combined into polymeric materials using a small number of simple reactions. Recently, Brocchini et al.[23,24] have reported one of the first examples of parallel synthesis of a small library of synthetic polymers using a combinatorial approach. One might expect that as combinatorial approaches begin to be applied more widely to polymer synthesis, or other materials for that matter, the need for computational methods for library design will emerge just as it has in pharmaceuticals and agrochemicals.

The objective of this study is to demonstrate that computational tools can be used to design synthetic polymer libraries for parallel synthesis using concepts now familiar in the life sciences such as molecular diversity, similarity, and quantitative structure−property relationships (QSPR). The polymer library of Brocchini et al.[23,24] was used to explore the application of stochastic cluster analysis[6] and genetic algorithm driven QSPR[25] to design diverse and focused libraries of copolymers. Instead of biological activity, the target properties are fundamental polymer properties such as glass transition temperature (Tg) and hydrophobicity as measured by the air−water contact angle (CA). Tg and CA are both important properties for assessing the potential of these polymers for use as degradable biomaterials. The Brocchini et al.[23,24] polymer library presents an ideal test case because it is a model of how combinatorial synthesis can be applied to synthetic polymers and because experimental data was reported for each member of the library.

Initially, I created a virtual library of the polymer repeat units that included all members of the set resulting from combination of two sets of monomers. This library was then converted to 2-D topological descriptions and submitted to analysis using the diversity algorithm in SimSearch[6] in order

* E-mail: Charles_Reynolds@rohmhaas.com.

10.1021/cc9900044 CCC: $18.00
Published on Web 06/05/1999

Table 1. Diphenol Monomer Structures

monomer	repeat	R	monomer	repeat	R
DTM	**1a**	methyl	DTiP	**1a**	*i*-propyl
DTE	**1a**	ethyl	DTiB	**1a**	*i*-butyl
DTB	**1a**	butyl	DTsB	**1a**	*s*-butyl
DTH	**1a**	hexyl	DTBn	**1a**	benzylic
DTO	**1a**	octyl	HTE	**1b**	ethyl
DTD	**1a**	dodecyl	HTH	**1b**	hexyl
DTG	**1a**	$-C_2H_4OC_2H_4OC_2H_4-$	HTO	**1b**	octyl

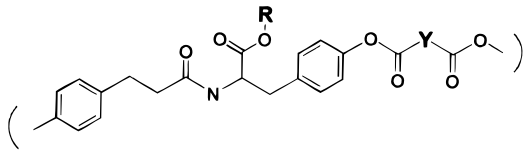
Table 2. Diacid Monomer Structures

monomer	diacid	Y
SCA	succinic	$-C_2H_4-$
GLA	glutaric	$-C_3H_6-$
DGA	diglycolic	$-CH_2OCH_2-$
AA	adipic	$-C_4H_8-$
MAA	3-methyl-adipic	$-CH_2CH_2CH(CH_3)CH_2-$
SUA	suberic	$-C_6H_{12}-$
DDA	dioxaoctane-dioic	$-CH_2OCH_2CH_2OCH_2-$
SBA	sebacic	$-C_8H_{16}-$

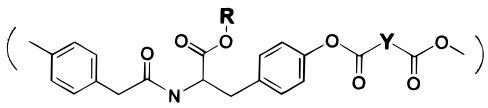
to identify a small subset of structurally diverse copolymers. This subset of diverse copolymers was then used to derive QSPR models for Tg and CA using genetic algorithm[25] (GA) selection of molecular descriptors. Once statistically significant models were derived, these models were tested both by predicting the properties of the rest of the polymer library and by using the QSPR models to design focused polymer libraries with specific combinations of properties (i.e. Tg and CA).

Computational Procedure

The basic copolymer repeat unit is given in structures **1a** and **1b**. The Cerius2[26] version 3.8 modeling package was used to enumerate a virtual library where R and Y are varied systematically as given in Tables 1 and 2. The structures are named using abbreviations suggested previously[23] for the diacid and diphenol monomers used to synthesize the copolymer. For example, the product of glutaric acid and



1a



1b

the methyl-substituted diphenol (**1a**, R = methyl, Y = C_2H_6) is denoted GLA-DTM. The monomer reagents are given in Tables 1 and 2 along with their abbreviations. The topology descriptors used in subsequent diversity selection and modeling of the copolymers were derived from the polymer repeat unit (**1a**, **1b**) capped with hydrogens. To avoid the additional complexity introduced by the conformational flexibility of the repeat units, molecular descriptors were only selected if they could be derived from the two-dimensional molecular connectivity. Previous work indicates that two-dimensional topology descriptors can be very useful in assessing molecular similarity or diversity,[27] and they can even be used to develop predictive quantitative structure−activity models for biological activity[28,29] or material properties.[30−32]

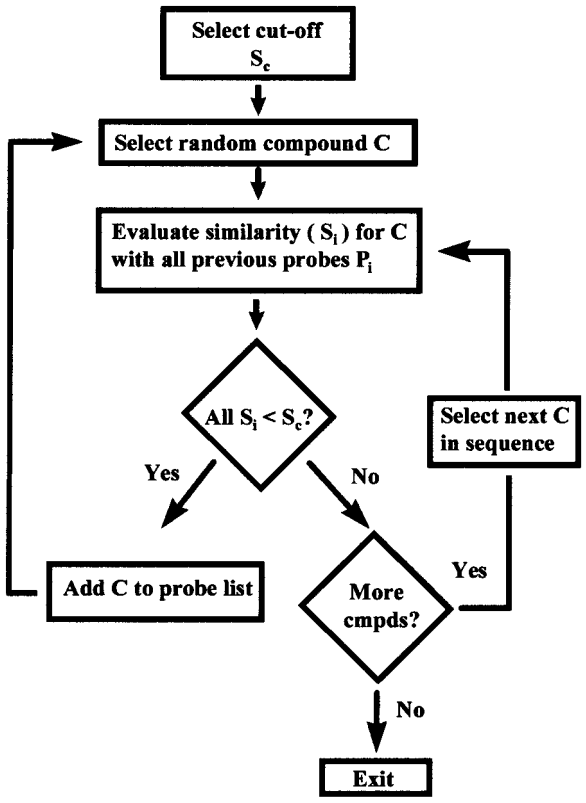


Figure 1. Schematic representation of the dissimilarity step in the stochastic cluster analysis algorithm implemented in SimSearch. S_c is the similarity cutoff used for assessing diversity, C represents any compound in the data set, and S_i is the similarity score with respect to one of the previous probes (P_i).

The stochastic search method implemented in SimSearch[6] was used to select a diverse subset of copolymer repeat units from the virtual library. SimSearch uses topology descriptors that are analogous to the topological torsion.[33] Each descriptor is divided into two parts: an atom type and topological path. In our implementation the atom type is defined by the element and the hybridization at that atom (e.g. C-sp^2, O-sp^3, etc.). SimSearch breaks the hydrogen-suppressed path for a given molecule into diads, triads, and tetrads (e.g. connected paths of 2, 3, and 4) of atom types representing all connected paths in the molecule. An example of this topology descriptor is given in the Supporting Information. The search algorithm for identifying diverse subsets of structural libraries uses a random search of descriptor space with a hard similarity cutoff for accepting compounds into the probe (diverse compound) list. This procedure has been published separately[6] and is only outlined briefly in Figure 1.

The QSPR models developed in this study were derived using the genetic algorithm[25] driven linear regression method in Cerius2. GA was used to guide variable selection for the least-squares fit. In each case 100 equations were selected randomly to begin the optimization. GA evolution was then used to vary the descriptors and number of descriptors in

Table 3. Diverse Polymer Sublibrary

diacid	diphenol	Tg	CA	diacid	diphenol	Tg	CA
SBA	DTD	12	101.2	DDA	DTD	17	94.4
DGA	DTBn	76	71.1	DDA	DTG	18	75.9
DGA	DTE	79	67.3	SCA	HTO	51	86.6
DGA	DTO	40	86.6	DGA	HTH	46	80.8
GLA	DTO	32	85.7	AA	HTE	65	71.4
AA	DTBn	61	74.1	DDA	HTE	50	67.7
AA	DTE	61	73.0	DDA	HTO	23	90.4
SUA	DTG	6	85.7	SBA	HTO	16	90.7
DDA	DTB	45	75.2				

the QSPR equation based on a lack-of-fit score.[34] The lack-of-fit metric in Cerius2 penalizes equations with poor sum-of-squares errors and larger numbers of parameters. Thus equations are favored that have maximal r^2 values and a minimal number of independent variables. In all cases at least two independent GA optimizations were run for 5000−10000 crossovers each. Multiple runs ensured that the results were converged and independent of the starting equations. I only employed linear terms in this study even though the program allows for quadratic and other nonlinear terms.

The topology descriptors used to derive the QSPR equations in this study include the standard Kier−Hall type descriptors,[35,36] log *P* calculated using ACD Labs software,[37] and other 2-D topology descriptors such as number of rotatable bonds and the Wiener index.[38] The complete set of descriptors employed in the GA optimization is given in the Supporting Information.

Selecting a Diverse Subset

A virtual library of 112 copolymers was generated based on the copolymerization of 14 diphenols with 8 diacids as described above (Tables 1, 2). The polymers in this library were further converted to SimSearch topology descriptors and submitted for diversity analysis using a similarity cutoff of 0.95. A high similarity cutoff was necessary because a significant portion of the polymer repeat units are conserved across the polymer library. The diversity search at 0.95 led to the selection of 17 polymers which are given in Table 3. It should be noted that this selection was based solely on the molecular structure and not any hypothetical model for Tg or CA. To assess how well the sublibrary represents the full polymer library with respect to Tg and CA, the experimental 2-D descriptor space defined by Tg and CA for all 112 copolymers is displayed graphically in Figure 2. Each member of the polymer library is represented as a point in Tg and CA space in Figure 2. The 17-member subset selected using diversity analysis is highlighted (filled circles). Visual inspection of Figure 2 shows that the sublibrary is actually quite diverse and provides good coverage of the experimental range of Tg and CA in the full library. The only exceptions are AA-DTBn and AA-DTE that give nearly identical Tg and CA values. However, the diversity algorithm might be forgiven for selecting both of these polymers since the side chains are quite different structurally with benzyl and ethyl substituents, respectively.

QSPR Models

The subset of 17 diverse polymers selected above was used to derive separate QSPR models for Tg and CA. These

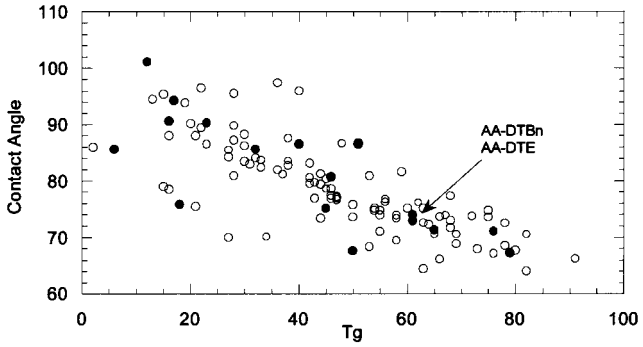


Figure 2. Plot of the property space for Tg and contact angle. The SimSearch selected subset is represented by filled circles. The rest of the 112-polymer library is represented as open circles.

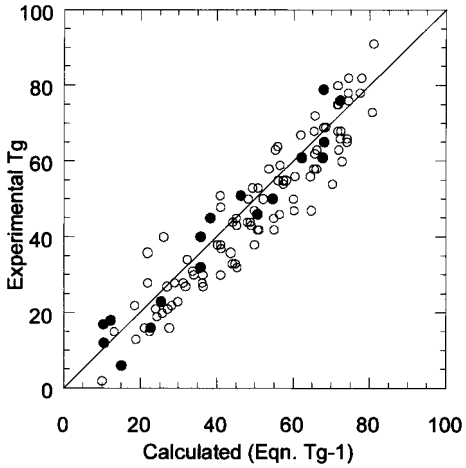


Figure 3. Calculated versus experimental Tg using model Tg-1. Diverse subset is represented by filled circles.

polymers represent approximately 15% of the total library for which experimental data are available. QSPR models were derived for Tg and CA, with respect to only these 17 copolymers, using GA driven linear regression as described above. As is typical, the GA optimization resulted in a collection of equations correlating various topology descriptors with the property of interest. Each GA generation contained 100 equations, so the final generation contained the most fit collection of 100 equations.

In the case of Tg, many equations were identified that had high r^2 values. A handful of descriptors such as number of rotatable bonds, Kappa-2, and Wiener index were represented repeatedly in these equations. To guard against overfitting the relatively small subset of polymers, I restricted consideration to equations containing three or fewer variables and an intercept. The most efficient models with an r^2 greater than 0.90 found in both independent GA runs each use only two descriptors. The descriptors found in these four models include number of rotatable bonds, Kappa-3, Kappa-2, Kappa-2-AM, and the Weiner index. Five representative models are given in Table 4.

The r^2 for the three-variable model, Tg-1, is 0.94. A plot of the calculated versus experimental Tg values is given in Figure 3. The data points used to derive this equation are shown as filled circles. Figure 3 illustrates the excellent correlation provided by this linear model, and further underlines the value of using diversity criteria to define the

Table 4. QSPR Models for Tg

	intercept	var_1	var_2	var_3	r^2	r^2 (all)[a]	RMS (all)[a] (°C)
Tg-1	202.588	−10.7389 (rot. bonds)	+0.009741 (Wiener)	+2.78809 (ACD log *P*)	0.94	0.89	7.1
Tg-2	181.365	−63.2841 (rot. bonds)	+77.3128 (Kappa-3)		0.93	0.59	23.7
Tg-3	151.08	−41.9474 (rot. bonds)	+44.8472 (Kappa-2-AM)		0.91	0.88	7.1
Tg-4	202.421	−10.6170 (rot. bonds)	+0.011611 (Wiener)		0.91	0.87	7.4
Tg-5	278.436	−16.9746 (Kappa-2)	+0.018791 (Wiener)		0.91	0.85	8.5
Tg-6	146.902	−4.67357 (rot. bonds)			0.82	0.83	8.3

[a] Error for the complete data set of 112 polymers using this model.

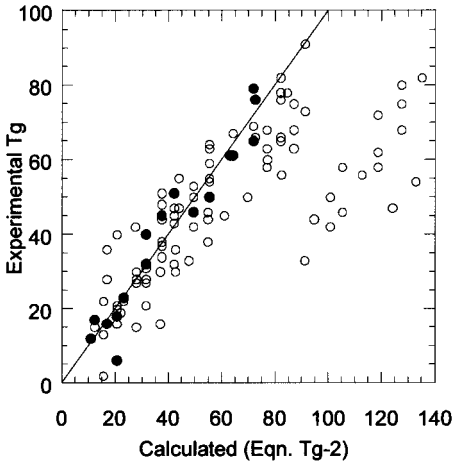


Figure 4. Calculated versus experimental Tg using model Tg-2. Diverse subset is represented by filled circles.

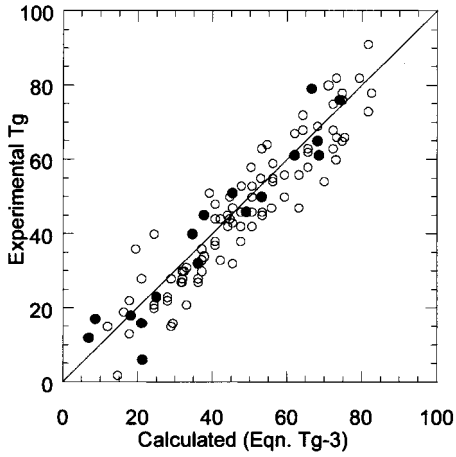


Figure 5. Calculated versus experimental Tg using model Tg-3. Diverse subset is represented by filled circles.

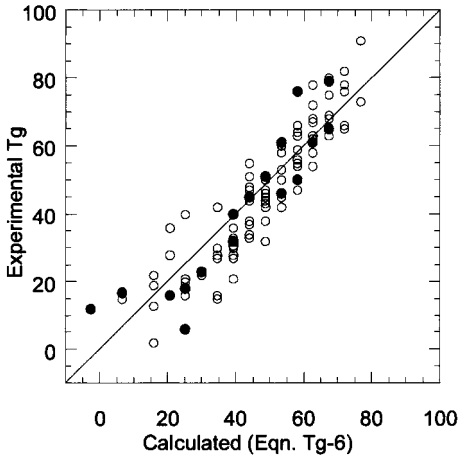


Figure 6. Calculated versus experimental Tg using model Tg-6 (number of rotatable bonds only). Diverse subset is represented by filled circles.

subset for modeling. The 17 copolymers give an evenly distributed range of Tg values that span the full range. There are four subset points clustered below a Tg of 20 °C, but those four points span a range of contact angles. A random selection of only 17 copolymers would run a significant risk of leaving part of the Tg range underrepresented. Figure 3 also shows that the Tg model derived from the subset of 17 copolymers is quite predictive for the other 95 copolymers that were not used to determine the model. The RMS error in calculated Tg for the full set is 7 °C. Similar comparisons of calculated and experimental Tg values show that the other simpler two-variable models are also quite predictive for the larger set of 112 polymers with the exception of equation Tg-2 (Figure 4). Model Tg-2 gives very poor predictions for all of the copolymers that involve the branched substituents such as DTiB and DTiP. These monomers are not represented in the subset of 17 copolymers and obviously lie on a separate line in Figure 4 relative to the other polymers. There is always a chance with any empirical model that it will fail for structures that fall outside the training set. In this regard, one advantage of the GA method is that it creates a population of equations for selection, not just a single equation. A plot of experimental versus calculated Tg is given for the two-variable model (Tg-3) in Figure 5. This represents an efficient model that is still very reasonable in terms of its predictive power.

Since the number of rotatable bonds shows up repeatedly in the GA optimization and a similar metric was proposed by Brocchini et al., a model that includes only the number of rotatable bonds is also given in Table 4. Even this simple model, based on a single variable, is remarkably good giving an r^2 of 0.82 (Figure 6) for the training set. It is intuitively reasonable that the Tg should be highly correlated with the number of rotatable bonds. It is well known that more flexible polymers tend to have lower glass transition temperatures. Consequently, any addition of rotatable bonds in the polymer backbone should lead to lower Tg values. In addition, longer more extended side chains in the diphenol monomer (R) should also lead to more rotatable bonds, greater free volume, and lower Tg. It is most likely that the other topology descriptors such as the Wiener index are capturing small effects due to branching in the diphenol side chain.

Analogous to the Tg models above, GA optimization was used to generate a set of equations for the air−water contact angle (CA). Again, I restricted consideration to equations containing three or fewer variables excluding the intercept. Representative equations that satisfy this requirement are

Table 5. QSPR Models for CA

eq	intercept	var_1	var_2	var_3	r^2	r^2 (all)[a]	RMS (all)[a] (deg)
CA-1	22.1744	+3.23187 (ACD log *P*)	+4.13635 (Kappa-2-AM)	−0.00534 (Wiener)	0.95	0.92	2.4
CA-2	58.2026	+7.64438 (CHI−V-1)	−6.93161 (CHI-3_P)		0.89	0.88	3.0
CA-3	65.6838	+7.64479 (CHI−V-1)	−1.75771 (SC-2)		0.89	0.88	2.9
CA-4	57.9674	+4.19735 (ACD log *P*)			0.82	0.87	3.0

[a] Error for the complete data set of 112 polymers using this model.

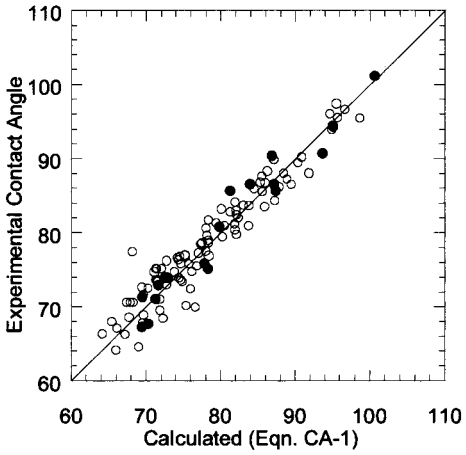


Figure 7. Calculated versus experimental CA using model CA-1. Diverse subset is represented by filled circles.

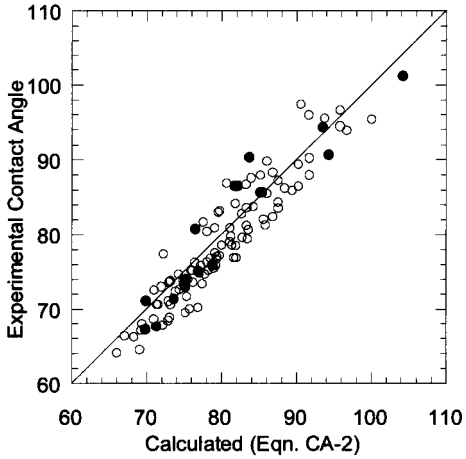


Figure 8. Calculated versus experimental CA using model CA-2. Diverse subset is represented by filled circles.

given in Table 5. In addition to the r^2 for the 17 polymers used in generating the model, the r^2 is also given for each model with respect to the full set of 112 polymers. Plots of the experimental versus calculated CA for the first two models, CA-1 and CA-2, are given in Figures 7 and 8, respectively. The three-variable model (CA-1) is excellent with an r^2 for the full 112 polymers of 0.92, but the two-variable model is only a little poorer with an r^2 of 0.88 (Figure 8).

The first 14 equations in one GA run and first 17 equations in the second GA run all contained ACD log *P* as an independent variable. This would seem to indicate an important role for this variable in modeling the contact angle. Indeed, an equation using only the ACD log *P* gives a respectable model with $r^2 = 0.82$ for the training set and $r^2 = 0.87$ including the polymers withheld from the training set (Table 5; Figure 9).

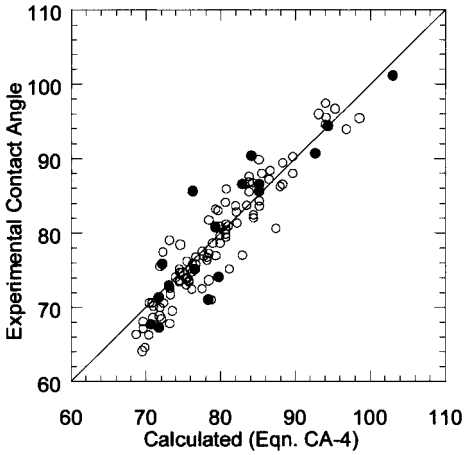


Figure 9. Calculated versus experimental CA using model CA-4. Diverse subset is represented by filled circles.

The models derived for both Tg and CA using the 17-polymer subset are predictive over the entire library at a level that would be very useful for polymer design. It might be possible to derive even better models using descriptors that have been designed more specifically for modeling polymers, such as those proposed by Bicerano.[39−41] For example, Bicerano makes a distinction between rotatable bonds in the polymer backbone and the side chain. This distinction might improve the model reported here for Tg. However, given the quality of the models derived from more standard topological descriptors, there was little justification for adopting more specialized descriptors.

Focused Library Design

The QSPR models can be used to assemble focused libraries that have a specific combination of desired properties. For example, if one wanted a library of hydrophobic polymers with low Tg, it would be possible to use the models developed previously for Tg and CA to score a virtual library of polymers and select only the polymers that are predicted to have the desired blend of properties.

The QSPR models for Tg and CA (Tg-1 and CA-1) were used to design two focused libraries: one (**Focus1**) with Tg between 0 and 20 °C and a contact angle between 80 and 100°, and a second (**Focus2**) with Tg between 60 and 80 °C and a contact angle between 60 and 80°. These libraries are given in Table 6 and shown graphically in Figure 10. Inspection of Figure 10 shows that the low Tg−high CA library has one member that falls slightly outside the design box and misses four compounds that fall within the design box. However, the "hit rate" of 55% is quite reasonable for a focused library. Further, if one wanted to be more certain of finding all of the compounds in the library that satisfy the design properties bounded by the box in the upper left

Table 6. Calculated and Experimental Results for Tg and Contact Angle

diacid	diphenol	Tg (expt), °C	Tg-1, °C	residual, °C	CA (expt), deg	CA-1, deg	residual, deg	focus[a]
SCA	DTB	67	62.0	5.0	74.0	74.4	−0.4	2
SCA	DTBn	78	77.7	0.3	72.6	70.3	2.3	2, 4
SCA	DTD	40	26.2	13.8	96.1	94.7	1.4	
SCA	DTE	78	74.5	3.5	68.7	67.8	0.9	2, 4
SCA	DTG	34	32.3	1.7	70.3	75.5	−5.2	
SCA	DTH	53	50.9	2.1	81.0	80.5	0.5	
SCA	DTM	91	81.2	9.8	66.4	64.2	2.2	4
SCA	DTO	48	41.1	6.9	86.8	86.0	0.8	
SCA	DTiB	75	71.9	3.1	74.8	71.1	3.7	2, 4
SCA	DTiP	82	78.0	4.0	70.7	67.9	2.8	2, 4
SCA	DTsB	75	71.6	3.4	73.7	71.3	2.4	2, 4
SBA	DTB	30	34.2	−4.2	83.6	85.9	−2.3	
SBA	DTBn	42	55.2	−13.2	80.7	78.1	2.6	
SBA	DTD	12	10.6	1.4	101.2	100.7	0.5	3
SBA	DTE	44	44.3	−0.3	79.5	80.2	−0.7	
SBA	DTG	2	10.1	−8.1	86.0	84.5	1.5	1, 3
SBA	DTH	20	25.8	−5.8	90.3	90.9	−0.6	
SBA	DTM	47	49.8	−2.8	77.3	77.1	0.2	
SBA	DTO	13	19.0	−6.0	94.6	95.0	−0.4	1, 3
SBA	DTiB	33	44.1	−11.1	82.5	82.1	0.4	
SBA	DTiP	44	48.9	−4.9	81.4	79.5	1.9	
SBA	DTsB	36	43.8	−7.8	82.1	82.3	−0.2	
DGA	DTB	64	56.0	8.0	72.5	76.0	−3.5	
DGA	DTBn	76	72.4	3.6	71.1	71.3	−0.2	2, 4
DGA	DTD	36	22.0	14.0	97.5	95.5	2.0	
DGA	DTE	79	68.1	10.9	67.3	69.5	−2.2	2, 4
DGA	DTG	27	27.1	−0.1	70.1	76.7	−6.6	
DGA	DTH	45	45.2	−0.2	80.5	82.0	−1.5	
DGA	DTM	82	74.6	7.4	64.2	66.0	−1.8	2, 4
DGA	DTO	40	35.9	4.1	86.6	87.2	−0.6	
DGA	DTiB	72	65.9	6.1	73.9	72.6	1.3	2, 4
DGA	DTiP	80	71.8	8.2	67.9	69.5	−1.6	2, 4
DGA	DTsB	68	65.6	2.4	73.1	72.8	0.3	2, 4
GLA	DTB	55	56.0	−1.0	74.9	76.2	−1.3	
GLA	DTBn	66	72.5	−6.5	73.8	71.5	2.3	2, 4
GLA	DTD	28	22.0	6.0	95.6	95.7	−0.1	
GLA	DTE	69	68.1	0.9	69.0	69.6	−0.6	2, 4
GLA	DTG	21	27.1	−6.1	75.6	76.9	−1.3	
GLA	DTH	43	45.3	−2.3	79.9	82.1	−2.2	
GLA	DTM	76	74.7	1.3	67.2	66.2	1.0	2, 4
GLA	DTO	32	36.0	−4.0	85.7	87.4	−1.7	
GLA	DTiB	62	65.9	−3.9	76.3	72.8	3.5	2, 4
GLA	DTiP	68	71.8	−3.8	71.8	69.7	2.1	2, 4
GLA	DTsB	58	65.6	−7.6	74.0	72.9	1.1	2, 4
AA	DTB	46	50.6	−4.6	77.6	78.1	−0.5	
AA	DTBn	61	67.9	−6.9	74.1	72.8	1.3	2, 4
AA	DTD	22	18.5	3.5	96.7	96.7	−0.0	1, 3
AA	DTE	61	62.3	−1.3	73.0	71.7	1.3	2, 4
AA	DTG	15	22.6	−7.6	79.1	78.4	0.7	3
AA	DTH	38	40.3	−2.3	83.7	83.8	−0.1	
AA	DTM	69	68.7	0.3	70.7	68.2	2.5	2, 4
AA	DTO	28	31.4	−3.4	87.3	88.9	−1.6	3
AA	DTiB	56	60.5	−4.5	76.9	74.6	2.3	2, 4
AA	DTiP	58	66.2	−8.2	73.5	71.5	2.0	2, 4
AA	DTsB	50	60.2	−10.2	75.9	74.8	1.1	2, 4
MAA	DTB	45	54.9	−9.9	78.7	77.6	1.1	
MAA	DTBn	60	72.9	−12.9	75.3	72.1	3.2	2, 4
MAA	DTD	19	24.4	−5.4	94.0	95.0	−1.0	
MAA	DTE	63	66.4	−3.4	75.3	71.5	3.8	2, 4
MAA	DTG	16	27.7	−11.7	78.6	77.3	1.3	
MAA	DTH	33	44.9	−11.9	83.8	83.1	0.7	
MAA	DTM	68	72.6	−4.6	77.5	68.2	9.3	2, 4
MAA	DTO	30	36.5	−6.5	86.3	87.8	−1.5	
MAA	DTiB	47	64.8	−17.8	76.8	74.2	2.6	2, 4
MAA	DTiP	54	70.4	−16.4	75.3	71.3	4.0	2, 4
MAA	DTsB	56	64.5	−8.5	76.4	74.4	2.0	2, 4
SUA	DTB	37	41.3	−4.3	81.3	81.9	−0.6	
SUA	DTBn	47	60.3	−13.3	77.1	75.3	1.8	2
SUA	DTD	15	13.2	1.8	95.5	98.7	−3.2	1, 3
SUA	DTE	50	52.2	−2.2	75.9	75.8	0.1	

Table 6. Continued

diacid	diphenol	Tg (expt), °C	Tg-1, °C	residual, °C	CA (expt), deg	CA-1, deg	residual, deg	focus[a]
SUA	DTG	6	15.1	−9.1	85.7	81.4	4.3	1, 3
SUA	DTH	27	31.8	−4.8	84.4	87.3	−2.9	
SUA	DTM	55	58.1	−3.1	74.1	72.5	1.6	
SUA	DTO	21	24.0	−3.0	88.1	91.9	−3.8	3
SUA	DTiB	42	51.1	−9.1	79.7	78.2	1.5	
SUA	DTiP	46	56.4	−10.4	77.0	75.3	1.7	
SUA	DTsB	46	50.8	−4.8	78.7	78.4	0.3	
DDA	DTB	45	38.5	6.5	75.2	78.4	−3.2	
DDA	DTBn	55	57.5	−2.5	71.1	71.9	−0.8	
DDA	DTD	17	10.5	6.5	94.4	95.2	−0.8	3
DDA	DTE	53	49.4	3.6	68.5	72.3	−3.8	
DDA	DTG	18	12.4	5.6	75.9	77.9	−2.0	1, 3
DDA	DTH	28	29.1	−1.1	81.0	83.8	−2.8	
DDA	DTM	63	55.4	7.6	64.6	69.0	−4.4	
DDA	DTO	16	21.2	−5.2	88.1	88.4	−0.3	3
DDA	DTiB	50	48.4	1.6	73.7	74.8	−1.1	
DDA	DTiP	58	53.7	4.3	69.6	71.9	−2.3	
DDA	DTsB	44	48.1	−4.1	73.5	74.9	−1.4	
SCA1	DTE	73	80.8	−7.8	68.1	65.4	2.7	2
SCA1	DTH	59	56.5	2.5	81.8	78.4	3.4	
SCA1	DTO	51	46.4	4.6	86.6	84.0	2.6	
DGA1	DTE	66	74.2	−8.2	66.3	67.2	−0.9	2
DGA1	DTH	46	50.7	−4.7	80.8	79.9	0.9	
DGA1	DTO	51	41.0	10.0	86.9	85.4	1.5	
AA1	DTE	65	68.2	−3.2	71.4	69.5	1.9	2
AA1	DTH	32	45.5	−13.5	84.2	81.9	2.3	
AA1	DTO	28	36.3	−8.3	89.9	87.1	2.8	
MAA1	DTE	63	72.0	−9.0	72.7	69.5	3.2	2
MAA1	DTH	38	50.0	−12.0	82.9	81.3	1.6	
MAA1	DTO	30	41.1	−11.1	88.4	86.2	2.2	
SUA1	DTE	54	57.5	−3.5	74.8	73.8	1.0	
SUA1	DTH	27	36.6	−9.6	85.6	85.6	0.0	
SUA1	DTO	22	28.3	−6.3	89.5	90.3	−0.8	
DDA1	DTE	50	54.8	−4.8	67.7	70.4	−2.7	
DDA1	DTH	31	33.8	−2.8	83.1	82.1	1.0	
DDA1	DTO	23	25.5	−2.5	90.4	86.9	3.5	
SBA1	DTE	43	49.1	−6.1	77.0	78.5	−1.5	
SBA1	DTH	23	30.0	−7.0	86.6	89.4	−2.8	
SBA1	DTO	16	22.8	−6.8	90.7	93.7	−3.0	
GLA1	DTE	65	74.2	−9.2	70.7	67.4	3.3	2
GLA1	DTH	42	50.7	−8.7	83.3	80.1	3.2	
GLA1	DTO	38	41.0	−3.0	87.7	85.5	2.2	
RMS error				7			2	

[a] Member of a focused library, e.g. 1 refers to **Focus1**.

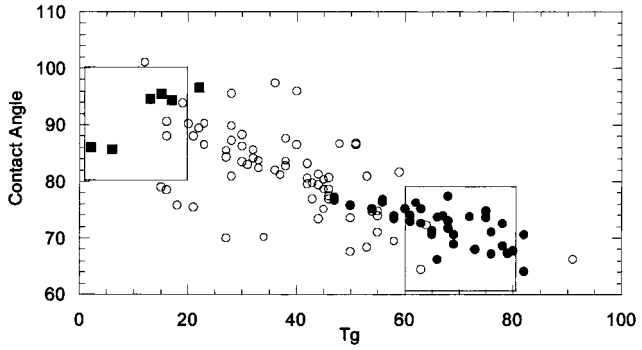


Figure 10. Map of Tg and CA property space. Two focused libraries are shown. The filled squares (**Focus1**) are a library focused on Tg between 0 and 20 °C and CA between 80 and 100° (low Tg−high CA). The filled circles (**Focus2**) are a library focused on Tg between 60 and 80 °C and CA between 60 and 80° (high Tg−low CA).

corner of Figure 10, one could loosen the selection criteria to include calculated values slightly outside the box. In the case of the low Tg−high CA library one would only need to increase the Tg cutoff from 20 to 24 °C in order to capture all of the polymers satisfying the target specifications.

Another way to measure the success of this focused library is to compute the number of "hits" identified for a given number of polymers synthesized and tested. For the low Tg−high CA focused library, 25 polymers (17 in the screening library + 8 in the focused library) nets five "hits". By comparison, random testing of 25 polymers from the full library of 112 possible polymers would only be expected to net two "hits" based on the incidence of polymers in the full library that satisfy both the Tg and CA restraints. Thus, using the screen and focus strategy provides more than twice as many polymers that satisfy our performance goals relative to random screening. Further, the strategy outlined above is likely to be much less variable than random screening where one might get lucky and find many polymers that satisfy the design criteria or, if less fortunate, find no polymers meeting the design criteria at all.

A second focused library (**Focus2**) was designed using the criteria of Tg between 60 and 80 °C and CA between

Table 7. Properties (°C) of Two Focused Libraries for Low Tg and High CA (**Focus1**, **Focus3**)[a]

	DTG		DTD		DTO	
	Tg	CA	Tg	CA	Tg	CA
SBA	2	86	12	101	13	95
DDA	18	76	17	94	16	88
SUA	6	86	15	96	21	88
AA	15	79	22	97	28	87

[a] Polymer compositions from library **Focus1** are enclosed in boxes. Other polymers are additions to fill out **Focus3**.

60 and 80°. This second high Tg−low CA library is shown as filled circles in Figure 10. In this case 26 of 28 compounds that satisfy this set of criteria were correctly identified using the QSPR equations to design the focused library. This amounts to a "hit rate" of 93%. This focused library also contains seven polymers that fall outside the defined property box, but only two of these are any significant distance from the desired range. In both cases the Tg values are poorly predicted.

Once again, it is useful to ask how well the screen and focus approach performs relative to random testing. In this case the screen and focus protocol would lead to synthesis and testing of 43 polymers (17 in the screening library + 8 in the focused library). This testing protocol provides 26 polymers with the desired Tg and CA values. By comparison random selection should net 13 polymers on average that have the desired properties. As was true with the first set of design criteria, the high Tg−low CA focused library provides approximately twice as many polymers meeting the design criteria as a random search even when the cost of making and testing the initial screening library is included. Of course, one could select any combination of Tg and CA for a focused library using the QSPR equations given in Tables 4 and 5. The two examples given here were simply chosen to illustrate the value of QSPR in designing focused polymer libraries.

Many times a significant issue when designing focused, or diverse, libraries is reagent efficiency.[42] Simply identifying the "best" compounds for synthesis, sometimes referred to as "cherry picking", may result in a library that is awkward to synthesize automatically and that is inefficient with respect to reagents. If reagent efficiency were a concern in this case, one could easily design alternative libraries where the monomers are input in row and column format. For example, in the first focused library (low Tg, high CA), one might simply select all reagents that lead to products in the selected subset (**Focus1**) and assemble a fully enumerated library based on these reagents (**Focus3**). This design (Table 7) doubles the number of polymers synthesized from 6 to 12, but leads to a filled block design and provides three additional hits. It actually provides five additional hits if one allows just ±1 °C of latitude in defining a hit. This library spans Tg and CA ranges of 2−28 °C and 76−101°, respectively. The largest deviation for a target property is 8 °C for Tg. This library is represented by filled squares in Figure 11.

In the case of the high Tg−low CA library (**Focus2**), there is a clear dichotomy between monomers that only appear once in the library and monomers that are present an average of six times in the library (Table 8). A second more efficient high Tg−low CA library was designed (**Focus4**) using only the monomers that occurred in more than one polymer in the initial library (**Focus2**). This leads to a block design library of 30 polymers derived from five diacids and six diphenols. This library is actually slightly smaller than **Focus2**, but maintains an impressive hit rate with only nine polymers falling outside the desired range. As was true for **Focus2**, **Focus4** falls entirely within the CA range, but with Tg values ranging from 47 to 91 °C, this library contains one polymer that falls 13 °C outside the desired range. This library is given as filled circles in Figure 11. Comparison of Figures 10 and 11 show that the focused libraries are of comparable quality regardless of whether they are selected purely based on the calculated properties of the product polymer (**Focus1**, **Focus2**) or if the libraries are recast so as to provide more experimentally efficient libraries (**Focus3**, **Focus4**) where the monomers are combined using a row and column format.

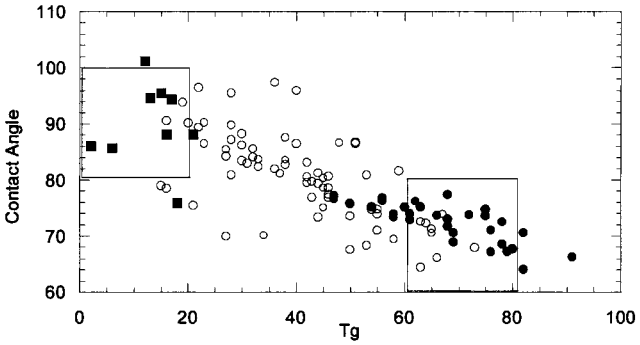


Figure 11. Map of Tg and CA property space. The reagent efficient focused libraries are shown. The filled squares (**Focus3**) are a library focused on Tg between 0 and 20 °C and CA between 80 and 100° (low Tg−high CA). The filled circles (**Focus4**) are a library focused on Tg between 60 and 80 °C and CA between 60 and 80° (high Tg−low CA). These libraries are comparable in quality to the "cherry picking" libraries in Figure 10.

Table 8. Incidence of Monomers Incorporated into the High Tg−Low CA Library (**Focus2**)[a]

diacid	incidence	diphenol	incidence
DGA	7	DTE	10
SCA	7	DTBn	6
AA	6	DTiB	5
GLA	6	DTiP	5
MAA	6	DTsB	5
		DTM	5
AA1, DGA1, GLA1, MAA1, SCA1, SUA	1	DTB	1

[a] **Focus4** is the complete 5 × 6 library of diacids DGA, SCA, AA, GLA, and MAA combined with the diphenols DTE, DTBn, DTiB, DTiP, DTsB, and DTM.

Conclusion

The same concepts of molecular similarity and diversity that have proven useful in the design of biologically active small molecules can also be applied in the design of synthetic polymers. Molecular topology descriptors and a stochastic diversity algorithm have been used to select a screening library of 17 polymers from a virtual library of 112 condensation polymers. Experimental data for the glass

transition temperature (Tg) and air−water contact angle (CA) for these 17 polymers have been used as a training set to derive a series of QSPR models for Tg and CA. These QSPR models have subsequently been used to calculate Tg and CA for the remaining polymers in the full virtual library. Comparison of the computed and experimental data show good agreement even for polymers that were withheld from the training set. Further, the QSPR models have been used to design focused libraries with specific target ranges of Tg and CA. It has been demonstrated that this screen and focus strategy is very effective for identifying polymers that fall in a specific range of Tg and CA. This general strategy should have widespread value as combinatorial design and synthesis are applied to the development of new synthetic polymers and materials.

Acknowledgment. I thank Professor Joachim Kohn and his research group at Rutgers University for their stimulating discussions and willingness to share their data.

Supporting Information Available. Complete set of descriptors employed in the diversity analysis and GA optimization. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Gordon, E. M.; Kerwin, J. F. *Combinatorial chemistry and molecular diversity in drug discovery*; John Wiley & Sons: New York, 1998.
(2) Chaiken, I. M.; Janda, K. D. *Molecular diversity and combinatorial chemistry. Libraries and drug discovery*; American Chemical Society: Washington, DC, 1996.
(3) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.
(4) Van Drie, J. H.; Lajiness, M. S. Approaches to virtual library design. *Drug Discovery Today* **1998**, *3*, 274−283.
(5) Warr, W. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.
(6) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(7) Cramer, R. D.; Patterson, D. E.; Clark, R. D.; Soltanshahi, F.; Lawless, M. S. Virtual Compound Libraries: A New Approach to Decision Making in Molecular Discovery Research. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010−1023.
(8) Johnson, M.; Lajiness, M.; Maggiora, G. Molecular similarity: a basis for designing drug screening programs. *Prog. Clin. Biol. Res.* **1989**, *291*, 167−71.
(9) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 339−353.
(10) Martin, E. J.; Spellmeyer, D. C.; Critchlow, R. E., Jr.; Blaney, J. M. Does combinatorial chemistry obviate computer-aided drug design? In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1997; Vol. 10, pp 75−100.
(11) Merritt, A. T. Uptake of new technology in lead optimization for drug discovery. *Drug Discovery Today* **1998**, *3*, 505−510.
(12) Hunt, D. A.; Baker, D. R.; Fenyes, J. G.; Basarab, G. S. Synthesis of agrochemicals and agricultural biotechnology entering the 21st century. In *Synthesis and Chemistry of Agrochemicals V*; American Chemical Society: Washington, DC, 1998; Vol. V, pp 1−5.
(13) Schultz, P. G.; Xiang, X.-D. Combinatorial approaches to materials science. *Curr. Opin. Solid State Mater. Sci.* **1998**, *3*, 153−158.
(14) Hsieh-Wilson, L. C.; Xiang, X.-D.; Schultz, P. G. Lessons from the Immune System: From Catalysis to Materials Science. *Acc. Chem. Res.* **1996**, *29*, 164−70.
(15) Senkan, S. M. High-throughput screening of solid-state catalyst libraries. *Nature (London)* **1998**, *394*, 350−353.
(16) Schlogl, R. Combinatorial chemistry in heterogeneous catalysis: a new scientific approach or "the king's new clothes". *Angew. Chem., Int. Ed.* **1998**, *37*, 2333−2336.
(17) Taylor, S. J.; Morken, J. P. Thermographic selection of effective catalysts from an encoded polymer-bound library. *Science (Washington, DC)* **1998**, *280*, 267−270.
(18) Wang, J.; Yoo, Y.; Gao, C.; Takeuchi, I.; Sun, X.; Chang, H.; Xiang, X.-D.; Schultz, P. G. Identification of a blue photoluminescent composite material from a combinatorial library. *Science (Washington, DC)* **1998**, *279*, 1712−1714.
(19) Xiang, X.-D.; Sun, X.; Briceno, G.; Lou, Y.; Wang, K.-A.; Chang, H.; Wallace-Freedman, W. G.; Chen, S.-W.; Schultz, P. G. A combinatorial approach to materials discovery. *Science (Washington, DC)* **1995**, *268*, 1738−40.
(20) Bratcher, M. S.; DeClue, M. S.; Grunnet-Jepsen, A.; Wright, D.; Smith, B. R.; Moerner, W. E.; Siegel, J. S. Synthesis of Bifunctional Photorefractive Polymers with Net Gain: Design Strategy Amenable to Combinatorial Optimization. *J. Am. Chem. Soc.* **1998**, *120*, 9680−9681.
(21) Reddington, E.; Sapienza, A.; Gurau, B.; Viswanathan, R.; Sarangapani, S.; Smotkin, E. S.; Mallouk, T. E. Combinatorial electrochemistry: a highly parallel, optical screening method for discovery of better electrocatalysts. *Science* **1998**, *280*, 1735−1737.
(22) Takeuchi, I.; Chang, H.; Gao, C.; Schultz, P. G.; Xiang, X.-D.; Sharma, R. P.; Downes, M. J.; Venkatesan, T. Combinatorial synthesis and evaluation of epitaxial ferroelectric device libraries. *Appl. Phys. Lett.* **1998**, *73*, 894−896.
(23) Brocchini, S.; James, K.; Tangpasuthadol, V.; Kohn, J. A Combinatorial Approach for Polymer Design. *J. Am. Chem. Soc.* **1997**, *119*, 4553−4554.
(24) Brocchini, S.; James, K.; Tangpasuthadol, V.; Kohn, J. Structure−property correlations in a combinatorial library of degradable biomaterials. *J. Biomed. Mater. Res.* **1998**, *42*, 66−75.
(25) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−66.
(26) *Cerius2, Version 3.8*; Molecular Simulations Inc., 9685 Scranton Rd., San Diego, CA 92121, 1998.
(27) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(28) Zheng, W.; Ethiraj, G. S.; Cho, S. J.; Tropsha, A. Comparative studies of 3D (CoMFA) and 2D QSAR methods. *Abstracts of Papers*, 215th National Meeting of the American Chemical Society, Dallas, March 29−April 2, American Chemical Society: Washington, DC, 1998; COMP-056.
(29) Zheng, W.; Cho, S. J.; Tropsha, A. Rational Combinatorial Library Design. 1. Focus-2D: A New Approach to the Design of Targeted Combinatorial Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.
(30) Hopfinger, A. J.; Koehler, M. G.; Rogers, D. Molecular modeling of polymers. 14. Quantitative structure−property relationship analyses of multicomponent systems containing polymers. *Macromol. Symp.* **1995**, *98*, 1087−100.
(31) Katritzky, A. R.; Sild, S.; Lobanov, V.; Karelson, M. Quantitative Structure−Property Relationship (QSPR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300−304.
(32) van Krevelen, D. W. *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*; Third Completely Revised Edition; van Krevelen, D. W., Ed.; Elsevier: Amsterdam, 1997; 898 pp.
(33) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−5.
(34) Friedman, J. Multivariate Adaptive Regression Splines. Laboratory for Computational Statistics, Department of Statistics, Stanford University 1988, p 102.
(35) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2, pp 367−422.
(36) *Molconn-Z*; eduSoft, LC, P.O. Box 1811, Ashland, VA 23298, 1996.
(37) *ACD LogP*, 1.0 ed.; ACD Labs, 141 Adelaide St., West, Suite 1501, Toronto, Ontario M5H 3L5, 1995.
(38) Canfield, E. R.; Robinson, R. W.; Rouvray, D. H. Determination of the Wiener molecular branching index for the general tree. *J. Comput. Chem.* **1985**, *6*, 598−609.

(39) Bicerano, J.; Editor. *Prediction of Polymer Properties*, 2nd ed., Revised and Expanded; Dekker: New York, 1996.
(40) Bicerano, J. Prediction of the properties of polymers from their structures. *J. Macromol. Sci., Rev. Macromol. Chem. Phys.* **1996**, *C36*, 161−96.
(41) Bicerano, J.; Sammler, R. L.; Carriere, C. J.; Seitz, J. T. Correlation between glass transition temperature and chain structure for randomly cross-linked high polymers. *J. Polym. Sci., Part B: Polym. Phys.* **1996**, *34*, 2247−2259.
(42) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

CC9900044