Designing Novel Polymers with Targeted Properties Using the Signature Molecular Descriptor

W. Michael Brown,*[,†,||] Shawn Martin,[†,||] Mark D. Rintoul,[†] and Jean-Loup Faulon[‡]

Department of Computational Biology, Sandia National Laboratories, P.O. Box 5800, Albuquerque, New Mexico 87185-0310, and Department of Computational Bioscience, Sandia National Laboratories, P.O. Box 969, Livermore, California 94551-9292

Received October 12, 2005

A method for solving the inverse quantitative structure−property relationship (QSPR) problem is presented which facilitates the design of novel polymers with targeted properties. Here, we demonstrate the efficacy of the approach using the targeted design of polymers exhibiting a desired glass transition temperature, heat capacity, and density. We present novel QSPRs based on the signature molecular descriptor capable of predicting glass transition temperature, heat capacity, density, molar volume, and cohesive energies of linear homopolymers with cross-validation squared correlation coefficients ranging between 0.81 and 0.95. Using these QSPRs, we show how the inverse problem can be solved to design poly(*N*-methyl hexamethylene sebacamide) despite the fact that the polymer was used not used in the training of this model.

INTRODUCTION

The ability to accurately design new molecules with targeted properties is a long-standing problem with potential applications in the design of catalysts, polymers, polymeric composites, solvents, detergents, drugs, pesticides, etc. At the heart of the problem, is the quantitative structure−property relationship (QSPR). In the forward-QSPR problem,[1] equations and algorithms which are capable of predicting molecular properties when given a structure are developed by empirical or theoretical means. In the inverse-QSPR problem, the molecular properties are the input; the output is the molecules which are calculated to satisfy the target properties. Together, the forward- and inverse-QSPR problems allow for computer-aided molecular design, which has the potential to greatly decrease the time and cost of trial-and-error procedures.

While the forward-QSPR problem has been a focus of intense research, literature presenting solutions to the inverse-problem is scarce. The most common solution to the problem is database search; a database of known compounds is evaluated using the forward-QSPR, and those most closely matching the desired properties are returned. The obvious limitation to this approach is that the search is restricted to known compounds present in the database. Perhaps more appealing is the ability to design novel compounds with desired properties. Approaches to such a design include random search,[2] combinatorial and heuristic-based enumeration,[3−6] graphical reconstruction,[7,8] mathematical programming,[9−20] and stochastic optimization.[1,21−27]

Each approach suffers from some disadvantage. Solutions to mixed integer nonlinear programming problems[9−20] are susceptible to local minima traps and are computationally expensive. Therefore, molecule templates or limitations on chemical groups for new molecules are commonly employed. Optimizations by simulated annealing, genetic algorithms, or tabu search[1,21−27] are random. While these algorithms are capable of reporting all near-optimal solutions in the search path, generating lists of new candidate molecules is subject to specifics of the search space and serendipity. This can be an important issue as candidate molecules often require further screening by molecular properties not present in the formulation (i.e. synthetic accessibility). The most straightforward and desirable solution would be an exhaustive enumeration of all candidate molecules within the chemical space.[3−6] Obviously, however, this approach is often infeasible due to combinatorial explosion.

We have previously presented an approach for the deterministic enumeration of molecules matching desired properties using the signature molecular descriptor.[28] In that work, we were able to design and validate novel potent inhibitors using an inverse-QSPR approach in which every molecule within the chemical space predicted to be a potent inhibitor was enumerated. However, in that case, the chemical space was given by cyclic peptides with a fixed size. Encoding was performed on the amino acid residue level such that all chemical groups had a valence of 2. For most chemical design problems, the combinatorial complexity will be much worse due to variable atom valences and variable molecule sizes.

Here, we demonstrate the efficacy of the approach for a more general (atomistic) design problem—the design of novel linear homopolymers. We show that the signature descriptor can be used in the forward-QSPR problem for prediction of several properties of interest to polymer design. We show that the inverse-QSPR problem can then be solved by limiting the chemical space to those polymers for which properties can be predicted without extrapolation. Finally, we validate the method with the rediscovery of the structure for Nylon-6,10.

* Corresponding author phone: (505)284-8938; fax: (505)844-5670; e-mail: wmbrown@sandia.gov.
† Department of Computational Biology.
‡ Department of Computational Bioscience.
|| These authors contributed equally to this paper.

10.1021/ci0504521 CCC: $33.50
Published on Web 01/28/2006

METHODOLOGY

Given an infinite chemical space, how do we find a polymer that will have the molecular properties desired for a particular application? Here, we present one approach to this problem, divided as follows:

1. The Chemical Space. Computations involving polymers require some vector space in which the molecular structure can be encoded.

2. The Forward-QSPR. To design novel polymers with desired properties, the ability to accurately predict those properties given a molecular structure is necessary.

3. The Inverse-QSPR. Once one is able to accurately predict a property given a structure, how does one solve the inverse problem of obtaining a molecular structure for a desired property?

4. The Confidence Metric. Given a large number of solutions to an inverse-QSPR problem, how does one choose which are best?

We present our solution to these issues, which are all interdependent, using the signature molecular descriptor.

The Chemical Space. Given the problem presented, the ideal molecular descriptor will allow an encoding such that both the QSPR can be efficiently calculated and the polymer structures corresponding to a given set of descriptors can be efficiently obtained. That is, the descriptors should give us both the property of interest and the molecules corresponding to that property. The signature descriptor is one such descriptor and is one of the few molecular descriptors for which a deterministic inverse-design has been shown to be efficacious.[28] The signature descriptor is designed to facilitate reconstruction of molecular graphs with a low degeneracy (small number of molecular structures corresponding to any given point in the chemical space) while at the same time providing a set of descriptors from which accurate QSPRs can be obtained for a wide variety of properties.[29]

Signature is based on the molecular graph of a molecule, $G = (V_G, E_G)$, where the elements in V_G denote the atoms in the molecule, and the edges of E_G correspond to the bonds between those atoms.[29] In the graph, both the nodes and the edges are labeled—the nodes by the element names of the atoms and the edges by the bond types (a single bond is unlabeled, a double bond is labeled by '=', a triple bond is labeled by 't', and an aromatic bond is labeled by 'p'). In this context, a molecule is characterized by a set of canonical subgraphs, each rooted on a different vertex with a predefined level of branching, which we refer to as the height h. The branching of a vertex is an extended degree sequence that describes the local neighborhood, up to a distance h away from the root.

We define an atomic signature, ${}^h\sigma_G(x)$, as the canonical subgraph of G consisting of all atoms a distance h from the root x. A molecular signature, ${}^h\Sigma_G$, is then the set of all unique atomic signatures and the number of times that they occur in the molecular graph. Even though the atomic signatures are unique, they are by construction interrelated, allowing information about the overall structure of the molecule to be recovered after analysis.

The atomic signatures make up the set of molecular descriptors for a molecule. These are expressed in terms of a string of characters that correspond to the canonized subgraph in a breath-first order. Branch levels are indicated by a set of parentheses following the parent vertex. For the case of linear homopolymers presented here, the molecular signature is calculated on a single polymer repeat unit. The atoms bonding repeat units are given a special label in the graph where the element name is followed by a subscript "z". An example of the height 1 molecular signature for poly(phenyl methylacrylate) is given in Figure 1. A more detailed explanation of signature has been previously published.[29]

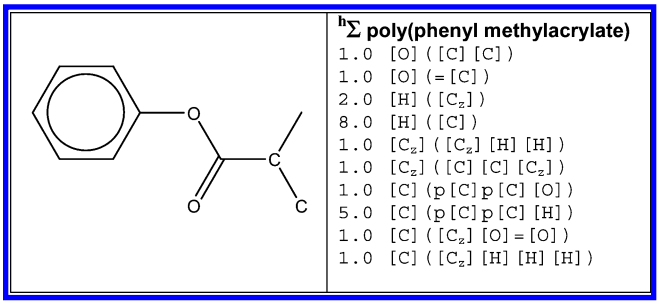


Figure 1. Molecular signature for poly(phenyl methylacrylate). The molecular signature on the right consists of two columns. The first column contains occurrence numbers, and the second column contains molecular fragments (atomic signatures). In this figure, the first row indicates that the fragment C−O−C occurs once in poly(phenyl methylacrylate), while the fourth row shows that H−C occurs eight times in poly(phenyl methylacrylate).

The Forward-QSPR. The forward-QSPR is the method of obtaining constraint equations in the polymer design process which restricts the entire space of all polymers to only those polymers with a desired property. Here we use multiple linear regression for obtaining QSPRs. Each of our QSPRs has an equation of the form $\Sigma\alpha_i x_i - \alpha_0 = P$, where α_i represents the regression coefficients, x_i represents the occurrence number of the molecular descriptor i, α_0 is the regression constant, and P is the property value of interest. When using the signature molecular descriptor to develop QSPRs, the signature height and number of descriptors must be chosen to optimize the predictive accuracy of the QSPR.

During the development of any QSPR, it is essential to avoid the often overlooked issue of overfitting.[30] When the number of descriptors is large relative to the number of molecules in the training set, it becomes easy to develop regressions based on chance correlations. When using signature descriptors, for example, large signature heights can often uniquely describe molecules. Although this can result in high correlation coefficients, the predictive accuracy of such models is likely to be very low. We therefore use the leave-one-out cross-validation squared correlation coefficient, a common statistical measure known as q^2. Using this approach, each polymer in the training set is predicted using a model that was not trained on the polymer in question, providing better statistics describing the ability of the model to accurately predict properties for novel polymers.

We have developed our QSPR models as follows. For each polymer in the training set, signatures were calculated at heights 0−7. Any signatures that occurred in less than 3 polymers were removed from descriptor selection as they were essentially constant across the data set. Signatures which were perfectly correlated were also removed from descriptor selection. QSPR equations were obtained using forward

Table 1: Prediction Accuracy for Polymer Properties Using the Signature Descriptor[a]

property	no. of mols in training set	no. of atomic signatures	optimal signature height	q^2	r^2	cross-validation absolute mean error	mean error % of range
T_g	261	132	1	0.81	0.93	27.97	5.77
C_p^s	51	12	0	0.91	0.93	16.62	4.27
ρ_a	99	92	1	0.86	0.89	0.05	4.54
E_{coh1}	87	13	0	0.95	0.95	4592.4	3.15
E_{coh2}	82	12	0	0.95	0.96	6037.3	2.35
V	98	12	0	0.93	0.94	7.40	2.76

[a] The squared correlation coefficients for cross-validation (q^2) and resubstitution (r^2) are given. The mean error percentages are calculated as the absolute mean errors divided by the range of experimentally determined values.

stepping feature selection as implemented in Matlab 7.[31] The routine stepwisefit.m was modified for efficiency and to prevent the addition of descriptors which would result in a QR factorization matrix that was close to singular. For optimization of the signature height and number of descriptors, q^2 was used as the objective function. All possible combinations of signature heights ranging from 0 to 7, and QSPR equation sizes (in terms of number of descriptors selected through forward selection), were evaluated, and the signature height resulting in the highest q^2 was reported. Optimization and cross-validation were also performed using Matlab scripts. The final QSPR model was trained using all polymers at the optimal signature height and descriptor count. QSPR models were trained on data compiled by Bicerano.[32] The number of polymers used for training is listed by property in Table 1. The full list of each polymer and corresponding property is included as Supporting Information.

The Inverse-QSPR. If we denote the space spanned by the signature descriptor to be the *signature space*, then we notice that not every point in the signature space corresponds to a possible polymer. For example, consider a point in the signature descriptor space consisting of a single atomic signature:

1.0 [C]([H][H][H][H])

This point does not correspond to a molecular signature because atomic signatures rooted at each atom are not present. In particular, a molecular signature containing the above atomic signature must also contain at least 4 occurrences of the atomic signature [H]([C]). Since our point in signature space does not contain 4.0 [H]([C]), it cannot correspond to a molecular signature.

To solve the inverse problem, we must therefore restrict the signature space. We use constraint equations (described in the next section) to achieve this restriction, giving us the *signature chemical space*, or just the chemical space. In the signature chemical space, we use the QSPR equation (forward-QSPR) to identify points that exhibit desired properties. These points are used to generate the molecular graphs corresponding to the identified molecular signatures. The end result is a list of polymers predicted by our method to have certain properties.

The Constraint Equations. The constraint equations enforce conditions necessary for reconstructing molecular graphs from points in the signature space. For the case of linear homopolymers, there are three types of constraint equations: the graphicality equation, the consistency equations, and the polymer repeat unit (PRU) equation.

The graphicality equation ensures that at least one connected graph can be constructed from the molecular descriptors. This equation, taken directly from graph theory, uses only the degree of the vertices in the graph. To build a connected graph, we require that (1) the sum of all the vertex degrees must be even and (2) the number of vertices of odd degree must be even. The resulting equation can be expressed in terms of a degree sequence $N = \{n_1, n_2, ..., n_k\}$ where n_i is the number of vertices of degree i. In this case, the degree sequence N is graphical if and only if there exists an integer $z \geq 0$ such that

$$\sum_{i=2}^{k}(i-2)n_i - n_1 + 2 = 2z \qquad (1)$$

The graphicality equation is a necessary condition for a graph to be connected and can be computed directly from the height zero molecular signature.

The next set of equations is collectively referred to as the consistency equations. Recall that a molecular signature is a collection of interrelated atomic signatures, where each atomic signature describes a particular atom and its neighboring atoms to a predetermined height. In constructing the signature of a molecule, it is guaranteed that a bond in one atomic signature will match up with a bond in another atomic signature, albeit in reverse order. However, blind reconstruction of the molecule requires equations to enforce these conditions of interdependency among the atomic signatures. This is done by matching bonds between two atoms of one signature to the bonds involving the same atoms in all other signatures.

We will use the notation ${}^h\sigma_i$ to describe the atomic signature of height h of an arbitrary atom i. Using ${}^h\sigma_i$ as a reference, any bond between the root and one of its children must be sought in all other atomic signatures in which the positions of the root and child are the transpose of ${}^h\sigma_i$. We use the notation $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j)$, to depict the number of bond types ${}^h\sigma_i$ has in common with ${}^h\sigma_j$. Clearly, then $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j) = \#({}^{h-1}\sigma_j \rightarrow {}^{h-1}\sigma_i)$. In the case where $i = j$, then $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_i)$ must be even. We note that the signature of a bond is one height less than the height of the molecular signature. When $\#({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j)$ is computed, one has to transpose the root i with a child j. While the neighborhood of i was initially probed up to height h, the transposed signature with new root j probes the neighborhood of j only up to height h-1.

The consistency equations can be summarized as follows: A molecular signature (${}^h\Sigma$) is consistent if and only if the two following conditions are verified:

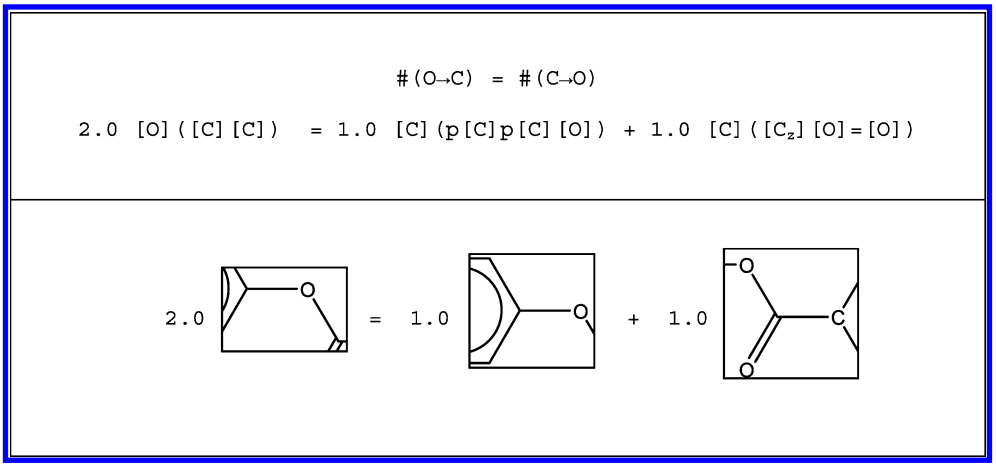


Figure 2. Consistency equation for poly(phenyl methylacrylate). Here we show the consistency equation for matching fragments containing O−C. This consistency equation describes mathematically the fact that O→C must occurs as many times as C→O in poly(phenyl methylacrylate). In this example, O→C occurs twice in the fragment [O]([C][C]), while C→O occurs once in [C](p[C]p[C][O]) and once more in [C]([C$_z$][O]=[O]).

a. For all atomic signatures ${}^h\sigma_i$ and ${}^h\sigma_j$ in ${}^h\Sigma$, # $({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_j) =$ # $({}^{h-1}\sigma_j \rightarrow {}^{h-1}\sigma_i)$.

b. For all ${}^h\sigma_i$ in ${}^h\Sigma$, # $({}^{h-1}\sigma_i \rightarrow {}^{h-1}\sigma_i)$ is even valued.

We give an example of a consistency equation for a height 1 signature in Figure 2. In addition, a more detailed explanation has been published previously.[28]

The final constraint equation is specific to the case of linear homopolymers and is necessary because the atoms bonding repeat units must be specified in order to lower the degeneracy of the signature descriptor; otherwise, the atoms joining repeat units within the polymer are unknown. Specifically, the constraint equation asserts that the sum of repeat unit bonding atoms, (labeled with a subscript '*z*') within any polymer must sum to exactly 2. This must be the case for a linear polymer.

Finding the Signatures for Novel Polymers. The constraint equations form a system of equations with unknown occurrence numbers x_i corresponding to atomic signatures *i*. The solutions to this system of equations are the molecular signatures for any polymers with the desired properties. These solutions must represent quantities meaningful to signature and hence, the occurrence numbers should take on non-negative integer values. Thus, our system of equations is Diophantine in nature. An algorithm adapted from Contejean and Devie[33] was implemented to solve our linear system of Diophantine equations. This algorithm uses a geometric interpretation of Fortenbacher's algorithm,[34] which efficiently solves homogeneous and nonhomogeneous linear Diophantine equations. Solution of the constraint equations gives us particular solutions to the nonhomogeneous equations as well as the Hilbert basis for the homogeneous equations. Using these solutions we can generate the signature chemical space. In combination with the QSPR equations, the molecular signatures for all possible polymers matching a given property can be obtained.

Our system of constraint equations is comprised of three types of linear equations: homogeneous, modulus, and nonhomogeneous. The consistency equations are comprised of homogeneous and modulus equations. However, the modulus equations can be rewritten as homogeneous equations by adding a dummy variable to enforce modularity. Thus our system effectively consists of homogeneous and nonhomogeneous equations. The graphicality constraint, the QSPR equations, and the PRU equation are nonhomogeneous. Of these, the QSPR is not usually Diophantine (does not have integer coefficients), and the graphicality equation is often too complex for the Diophantine solver. Therefore, we include in the constraint equations only the homogeneous (including converted modulus) equations and the PRU equation; the QSPR and graphicality constraints are enforced based on postprocessing. To enforce the PRU equation, we obtain particular solutions to the constraint equations (homogeneous and PRU). These particular solutions are added to a Hilbert basis obtained by solving the constraint equations where the PRU equation is forced to zero.

Chemically, linear combinations of the Hilbert basis vectors generate solutions without PRU bonding atoms. When those combinations are added to the particular solutions, signatures for complete molecules are obtained, such that these molecules include repeat unit bonding atoms. Thus the signature chemical space is spanned by the particular solutions (to the homogeneous equations and PRU equation) plus all linear combinations of the Hilbert basis vectors (general solutions). Finally, these solutions are screened using the graphicality equation and the QSPR equations to eliminate unfeasible or uninteresting possibilities.

Polymer Construction. Once all molecular signatures predicted to have the desired properties have been enumerated, the actual polymers must be found by reconstructing the molecular graphs from the signatures. The algorithm that enumerates all molecular graphs corresponding to a target signature is based on an isomer enumeration algorithm published some time ago.[35] This algorithm was recently adapted to enumerate isomers matching user specified signatures.[36] Starting with a molecular graph containing all atoms but no bonds, the algorithm belongs to the class of orderly algorithms,[37] where bonds are added in all possible ways in order to produce all nonisomorphic saturated graphs matching the target signature. The original idea of this algorithm is to saturate all equivalent atoms at once. Specifically, the atoms are first partitioned into equivalent classes, the classes being the orbits of the automorphism group of the graph, then an orbit is chosen, and all the possible graphs that can be generated by saturating all atoms

of the orbit are generated. The initial atom partitioning is performed following the target signatures. The process is recursive until all orbits have been saturated.

The number of molecules which can be reconstructed from a given target signature is the degeneracy of that signature. Ideally, the degeneracy of a descriptor should be low as low as possible while still allowing for high correlation with molecular properties.[38] For signature, the degeneracy can be decreased by increasing the signature height;[36] however, at large signature heights most correlations result from overfitting. Therefore, in practice we find that we have to sacrifice some degeneracy in order to obtain predicative accuracy in the QSPR equations. A detailed description of the reconstruction algorithm, and a study on descriptor degeneracy can be found in ref 36.

The Confidence Metric. In theory, the methods presented above allow one to perform an exhaustive enumeration of all possible polymers predicted to have desired properties. In practice, however, such an enumeration is unreasonable not only because it is too costly to compute but also because the domain under which the QSPR can be expected to be accurate is limited. A much more feasible (and desirable) objective is to enumerate all polymers for which the property calculation does not result from an extrapolation. In other words, we want to enumerate those polymers for which we have the highest confidence that the property prediction was accurate. We do this by imposing several limitations on which polymers will be considered for enumeration.

First, we consider only those signatures present in the property training set and therefore consider only a subset of the chemical space. This is justified since it is clearly an extrapolation to perform property prediction on polymers with chemical groups unknown to the forward-QSPR. This limitation is easily implemented by generating constraint equations using only the atomic signatures which are present in the training set.

Next, we take measures to prevent property extrapolation on polymers which are significantly different in size from the molecules in the training set. We impose this limitation in two forms. First, we limit the maximum occurrence of any atomic signature in a potential polymer based on the maximum occurrence of that signature in the training set. Second, we impose limitations on the number of basis vectors with nonzero coefficients in any linear combination and also limitations of the values for the coefficients of those nonzero coefficients.

The limitation on the number of nonzero basis vectors limits the size of potential polymers because our Hilbert basis and coefficients are restricted to non-negative integers. Therefore, there is an increase in the average number of atoms per polymer associated with an increase in the number of basis vectors used within a linear combination. More importantly, this limitation must be imposed simply because in many cases it is computationally infeasible to consider all linear combinations of all basis vectors. For n possible coefficients for each basis vector and b basis vectors, n^b possible linear combinations exist. Therefore, if the target molecular weight of the enumerated molecules is high and the number of basis vectors is high, an exhaustive search of the chemical space can be unattainable.

The coefficients used in the linear combinations are limited based on a target atom count (TAC) restriction which is a count of the number of atoms within a polymer. The limitation is implemented such that if the TAC is 50, every possible polymer containing up to 50 atoms is considered for enumeration. This is done by computing the maximum value of a given coefficient in a linear combination assuming that the other coefficients are one, and then considering all combinations of coefficients, up to the maximum computed value for each coefficient. Note that by asserting that all possible polymers up to a certain atom count are considered, we also must enumerate some polymers with atom counts greater than the TAC when using this method.

The final limitation used to restrict the generation of novel polymers is a confidence metric. While the above limitations simply ignore polymers with chemical groups or sizes unknown to the QSPR (which are obvious extrapolations), the confidence metric assigns a scalar value to each potential molecular signature, which is used to rank polymers by the expected prediction accuracy for a given property. This brings up the critical question important to all machine learning models: "How do we know which predictions are accurate and which predictions are inaccurate?". This question has recently been addressed in the cheminformatics arena with the development of discriminators for prediction accuracy in QSPRs.[39−41] Using this type of approach, one can determine how much trust to put into a given prediction. As an example, in our recent work involving a classification for protein structure prediction,[42] we had an overall prediction accuracy of 51.3% for an ordering problem. However, by using a confidence metric inherent to the model, we were able to achieve over 95% accuracy on the 25% of the data set with the highest confidence. The ability to perform such a discrimination is particularly relevant here, where one must sort through potentially large numbers of enumerated polymers matching given properties.

For this work, we have chosen to use a metric reported by Sheridan et al.[41] In a large study utilizing data sets covering 8 different properties and tens of thousands of training molecules, they found that the best-predicted molecules were those with the highest similarity and/or the most neighbors in the training set. This result is intuitive in that we expect regressions to have the highest accuracy in domains nearest the training data. In some cases, the authors went so far as to produce error bars for each prediction of a given property. Therefore, for this work, we have chosen to use a confidence metric which is calculated as the normalized Euclidean distance to the nearest molecule in the training set. (To calculate the normalized Euclidean distance, we first scale each molecule to have unit norm in the signature chemical space and then compute the Euclidean distance between the two molecules.)

Using the confidence limitations presented, we can reduce the run-time of the algorithm by limiting the chemical space and the number of basis vectors. We can reduce the number of enumerated polymers using maximum occurrence limitations and a confidence metric cutoff. Instead of arbitrarily choosing some portion of the chemical space to enumerate, we isolate the subset where we have the highest confidence in our predictions.

SUMMARY

The inverse-QSPR methodology presented herein offers a procedure for searching through the entire chemical space

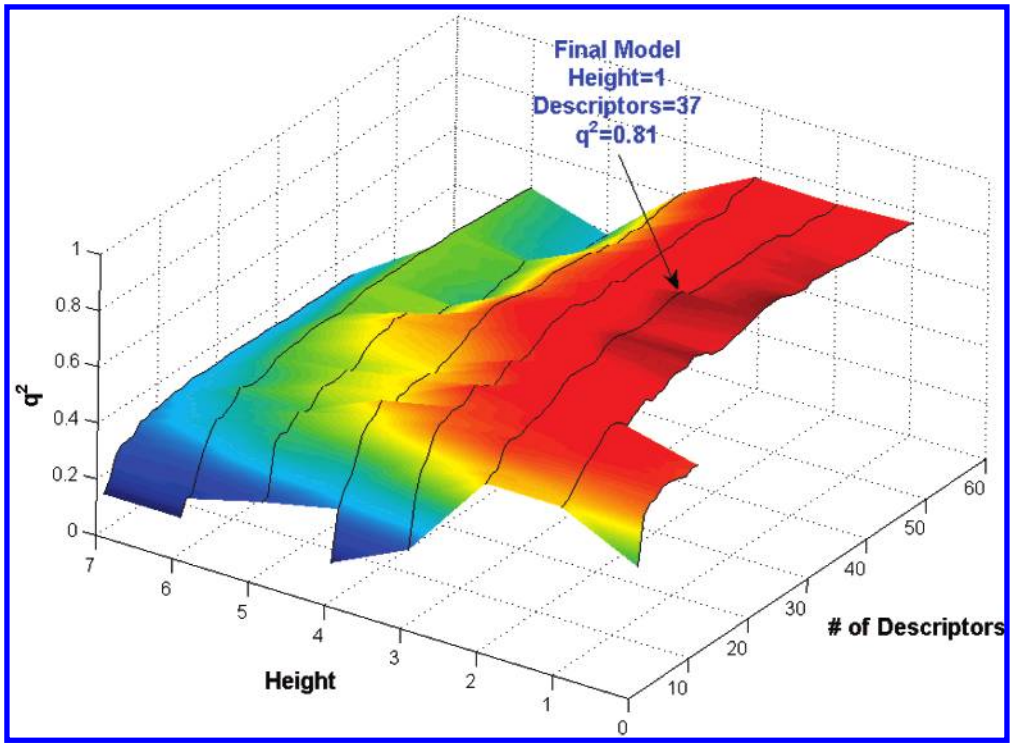


Figure 3. Optimization of the signature height for a set of linear homopolymers with glass transition temperature data. The optimum signature height is selected as the one which produces the highest cross-validation squared correlation coefficient in order to preserve predictive accuracy.

of polymers to isolate those matching desired properties in a deterministic manner. For certain applications, it is possible to enumerate every possible polymer within a given confidence threshold. In cases where the signature chemical space given by the training set is too large, or where the molecular weights of the training set are high, such an exhaustive search may not be possible. However, it is always that case that at least some subspace can be searched thoroughly to produce polymers with high confidence predictions.

To demonstrate the efficacy of the approach, we first show that the signature descriptor can be used in a forward-QSPR approach to generate equations for predicting certain properties of interest to polymer design. We then show, similar to the approach of Camarda and Maranas,[9] that these QSPRs can be used in the inverse design process to design a target polymer for which the properties have been experimentally determined but were used nowhere in the training of the model.

RESULTS

Predictive QSPRs are essential for the accurate design of novel polymers, and it is therefore important to assess the predictive accuracy of regression models developed using the signature descriptor. Here, we have chosen 6 properties of interest to polymer design—the molar volume (V), the amorphous density (ρ_a), the Fedors-type cohesive energy (E_{coh1}), the van Krevelen-type cohesive energy (E_{coh2}), the specific heat capacity (C_p^s), and the glass transition temperature (T_g). In developing QSPRs for these properties, there are two important issues to consider. First, the appropriate signature height must be chosen, and second, care must be taken to avoid overfitting the QSPR. To address both issues, we use a brute-force evaluation of cross-validation squared correlation coefficients (q^2) at each signature height between 0 and 7, and each possible number of descriptors as selected by forward stepping. The final model results from training on the entire data set using the signature height and descriptor count assessed to have the highest predictive accuracy as measured by the q^2. A representative plot showing q^2 as a function of height and descriptor count is shown in Figure 3 for the T_g data set. A summary of results for the 6 data sets is shown in Table 1. At worst, the mean error encompasses under 6% of the range for a given property. The q^2 ranges from 0.81 to 0.95, and the r^2 (for the final models) ranges from 0.89 to 0.96. Cross-validation plots of experimental versus predicted values for the 6 properties are given in Figure 4.

For the design problem, we decided to search for polymers similar to Nylon-6,10, which has a T_g of 313 K, a C_p^S of 439 $J \cdot mol^{-1} \cdot K^{-1}$, and a ρ_a of 1.04 $g \cdot cm^{-3}$. To avoid extrapolation, three constraints were enforced during molecular construction. First, the signatures and constraint equations used for polymer reconstruction were calculated using only the 33 polymers for which data were available for all three properties (see Supporting Information). Second, the maximum occurrence of any signature within a new polymer was limited to the maximum occurrence of that signature within the 33 polymer reconstruction data set. Third, the signature molecular descriptor for any new polymer was restricted to lie within a normalized Euclidean distance of 0.1 from the molecular descriptors in the reconstruction data set. Using this approach, we limit the design of new polymers to produce a manageable number of polymers in regions of the chemical space where we expect prediction accuracy to be highest.

Signature calculation for the 33 polymers resulted in 63 unique atomic signatures from which 23 homogeneous equations, 3 modulus equations, 1 graphicality constraint,

Table 2: Results from Signature Enumeration[a]

no. of homogeneous basis vectors	no. of possible signatures	signatures satisfying max. occ.	signatures within distance	acceptable signatures	mean no. of atoms per molecule	no. of training molecules recovered
0	1286	615	12	12	8.1	12
1	375 431	155 042	115	46	27.5	7
2	27 514 763	9 871 753	1130	346	41.2	9
3	782 388 775	240 705 584	3315	923	45.5	4
Total	810 280 255	250 732 994	4572	1327	43.4	32

[a] Each signature represents a point in the chemical space from which molecular graphs can be reconstructed to generate polymers. The mean number of atoms per molecule will go up with the number of nonzero basis vectors included in the linear combinations used for enumeration.

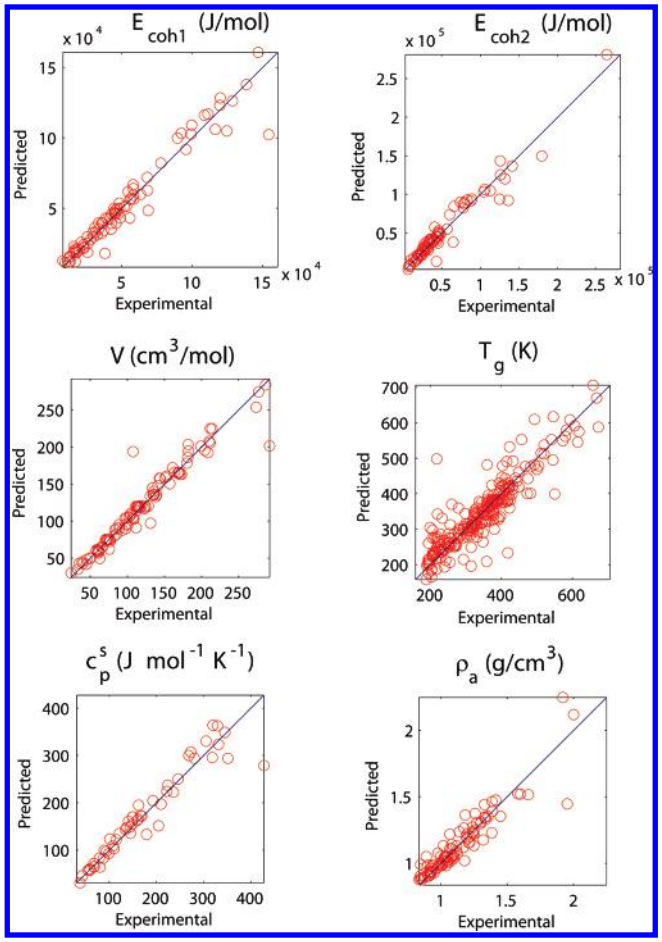


Figure 4. Cross-validation plots of predicted versus experimental values for several polymer properties. The value of the property for each point plotted is predicted using only the other data points for training.

and 1 PRU constraint were derived. Solutions to the homogeneous system resulted in a Hilbert basis of 25 vectors. The basis for the nonhomogeneous system gave 4707 particular solutions. Removing basis vectors that violated the maximum occurrence constraint left 19 homogeneous basis vectors and 2907 particular solutions. Molecular signature enumeration was performed using all linear combinations of up to 3 basis vectors and one particular solution, with the TAC set such that no polymer with under 50 atoms would be missed. The results from the enumeration, separated by the number of homogeneous basis vectors with nonzero coefficients, are listed in Table 2. Restriction of the basis vector coefficients and the signature space dimensionality based on the 33 polymer reconstruction data set resulted in a solution space of over 800 million polymer signatures. Restriction of this space by the maximum occurrences reduced the number by about 69%, while restriction of this space based on the confidence metric reduces the number by over 99%. When all constraints are considered, we found only 1327 polymer signatures which, by our measures, are unlikely to represent extrapolations from the model.

Due to combinatorial explosion, it is intractable to consider polymers whose signatures result from linear combinations where all basis vectors have nonzero coefficients. Here, we have made a restriction where only 3 basis vectors (and one particular solution) were allowed nonzero coefficients. How much did we miss? Based on the 50 atom limit, the entire solution space is comprised of over 800 million polymer signatures. We have considered less than 0.0002% of these potential solutions. However, with an increase in the number of basis vectors, there is also an increase in the mean number of atoms within the solutions. As is indicated in Table 2, when 3 basis were used, the average number of atoms per polymer is 45.5, while the average for the reconstruction data set is only 18.9. This indicates that we are largely uninterested in over 99% of the solutions that we did not calculate, just based on molecule size alone.

Ignoring any molecular property constraints, 1327 molecular signatures were obtained with molecular property distributions as shown in the histograms in Figure 5. All except one out of the 33 polymers used for training were reconstructed. The exception, bisphenol A polycarbonate, had the third highest molecular weight of the group and required a linear combination of over 3 basis vectors to represent its signature. Finally, we screened the 1327 molecular signatures to find solutions matching the target properties. This screen was performed by using a range around the target properties (313 K for T_g, 439 J·mol^{-1}·K^{-1} for C_p^s, and 1.04 g·cm^{-3} for ρ_a) of plus or minus their respective absolute mean errors as computed using the QSPRs (see Table 1). In the end, we obtained 80 signature molecular descriptors predicted to exhibit the desired properties.

Reconstruction of the 80 molecular descriptors resulted in 178 polymers predicted to be within the target range. Eight examples of the enumerated polymers are given in Table 3. Most importantly, our method enumerated the structure for Nylon-6,10 (Table 4) which has a 313K T_g, a 439 J·mol^{-1}·K^{-1} heat capacity, and a 1.04 g·cm^{-3} density. This validates our approach as the polymer was used nowhere in the training of this model. The entire process (signature calculation, QSPR calculation, constraint generation, basis calculation, reconstruction, etc.) required less than 2 days of real-time computation on a single processor desktop PC.

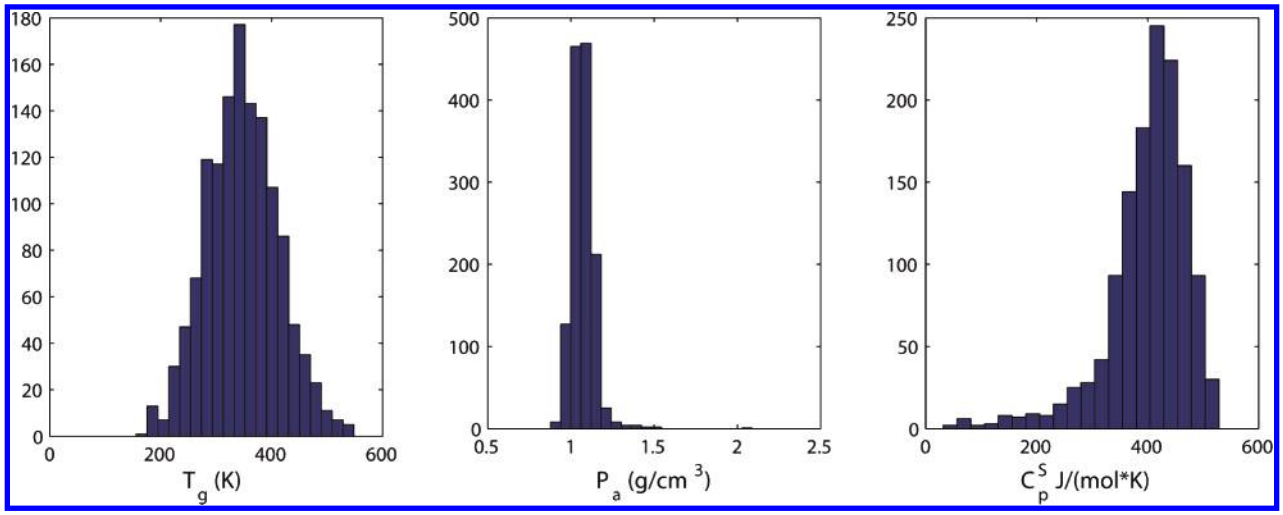
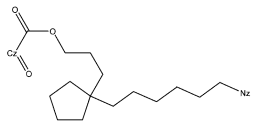
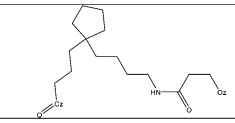
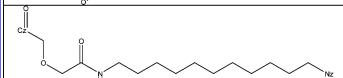
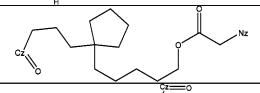
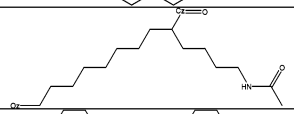
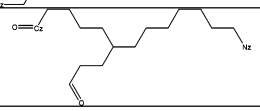
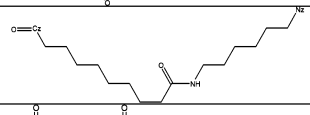
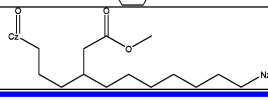


Figure 5. Histograms illustrating the QSPR calculated values for enumerated polymers using the inverse-design procedure with no targeted values.

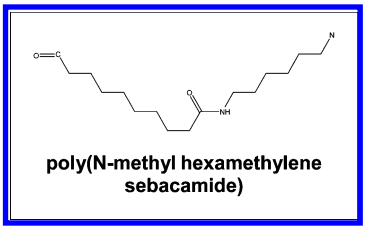
Table 3: Eight of the 178 Polymer Repeat Units Generated from Inverse-QSPR with Properties Targeted for a T_g of 313 K, a C_p^s of 439 J·mol^{-1}·K^{-1}, and a ρ_a of 1.04 g·cm^{-3}

Structure	T_g (K)	C_p^S (J·mol^{-1}·K^{-1})	ρ_a (g·cm^{-3})
	308.2	437.8	1.06
	319.5	434.4	1.06
	305.4	449.0	1.08
	299.9	437.8	1.06
	324.5	448.4	1.03
	295.8	432.7	1.00
	327.7	455.4	1.05
	294.1	426.3	1.04

DISCUSSION

The ability to perform in silico design of novel molecules with desired properties has the potential to greatly decrease both the time and cost of trial-and-error procedures. A deterministic enumeration of lists of compounds satisfying the target properties is advantageous in that it allows further processing based on properties which are difficult to include in the inverse-design formalism. For example, it is infeasible to include properties which require long computation times in an inverse-design approach. Additionally, properties such as synthetic accessibility are inherently difficult to represent mathematically and optimization to produce a single molecule is likely to produce one that is difficult to synthesize.

Table 4: Structure for Nylon-6,10 Generated from a Targeted Design of Polymers with a T_g of 313 K, a C_p^s of 439 J·mol^{-1}·K^{-1}, and a ρ_a of 1.04 g·cm^{-3} [a]



T_g (K)		C_p^s (J·mol^{-1}·K^{-1})		ρ_a (g/cm^3)	
exp	pred	exp	pred	exp	pred
313	327.7	439	455.4	1.04	1.047

[a] The polymer was used nowhere in the training of the model. Experimentally determined and QSPR-calculated properties are given.

Despite these advantages, deterministic enumeration for computer-aided molecular design is often dismissed due to the combinatorial complexity of the problem. Indeed, the chemical space is potentially infinite, and, therefore, an exhaustive consideration of all possible molecules is seemingly impossible. An important observation, however, is that most forward-QSPR formulations are empirical. There is a limited domain under which the predictions are accurate, and therefore only a subset of the chemical space should be considered for enumeration. Using this observation, we were able to limit the entire chemical space of linear homopolymers to ~800 million by throwing out compounds with a large number of atoms and descriptors which did not occur in the QSPR training sets. Further restrictions based on maximum descriptor values and similarity to molecules in the training set resulted in an enumeration of 178 molecules out of the chemical space for which we expected accurate predictions of desired properties.

We have demonstrated the feasibility of deterministic enumeration for inverse-design with our ability to rediscover 32 of the 33 polymers in the reconstruction data set and the ability to design a molecule novel to the model with experimentally determined properties. The ability to perform

inverse-design using the signature descriptor is complemented by the ability to develop correlations for several properties of interest to polymer design, and, therefore, use of the signature descriptor for forward-QSPR and inverse-QSPR is a promising approach for the computer-aided molecular design of novel polymers. In this work, we have considered the design problem only for linear homopolymers. While the extension to other polymeric macromolecules (branched, cross-linked, etc.) might be straightforward, future work is required to demonstrate feasibility in these cases.

There are many properties of interest to polymer design that we have not considered herein. In our experience, the types of properties that can be considered for signature inverse-design is limited, as with all machine-learning approaches, by the amount of experimental data available for training. One example for which we could not obtain an accurate QSPR was the prediction of polymer oxygen permeability. While a regression has been reported with high correlation between topological indices and oxygen permeability,[32] no cross-validation was performed in order to provide a valid assessment of predictive accuracy. Using signature, we were able to develop regressions with high correlation; however, the cross-validation squared correlation coefficients were low, presumably due to the low number of experimental data points available for training (57).

In cases where accurate QSPRs cannot be obtained using signature, potentially any descriptor can be used to further screen compounds in the inverse design process. The caveat, however, is that the use of other descriptors makes it difficult to limit the chemical space considered for enumeration during inverse design. That is, it may prove difficult to generate constraint equations in terms of signature (which is currently required for our inverse-design) which limit the chemical space in order to prevent extrapolation on QSPRs which are not expressed in terms of signature. Therefore, the use of at least *some* signature QSPRs is necessary in order to rationally limit the chemical space and the combinatorial explosion during our inverse-design process.

Software for signature calculation is available for free at http://www.cs.sandia.gov/~jfaulon/QSAR/index.html.

ACKNOWLEDGMENT

Funding for this work was provided by the U.S. Department of Energy and Sandia National Laboratories under Grant number DE-AC04-76DP00789. Sandia is a multiprogram laboratory operated by Sandia Corporation, a LockheedMartin Company, for the United States Department of Energy's National Nuclear Security Administration.

Supporting Information Available: Table of all experimental data used for QSPR training in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1994**, *35*, 188−195.

(2) Derringer, G. C.; Markham, R. L. A Computer-Based Methodology for Matching Polymer Structure with Required Properties. *J. Appl. Polym. Sci.* **1985**, *30*, 4609−4617.

(3) Brignole, E. A.; Bottlini, S.; Gani, R. A Strategy for Design and Selection of Solvents for Separation Processes. *Fluid Phase Equil.* **1986**, *29*, 125−132.

(4) Gani, R.; Nielsen, B.; Fredenslund, A. A Group Contribution Approach to Computer-Aided Molecular Design. *AIChE J.* **1991**, *37* (9), 1318−1332.

(5) Gani, R.; Brignole, E. A. Molecular Design of Solvents for Liquid Extraction Based on UNIFAC. *Fluid Phase Equil.* **1983**, *13*, 331−340.

(6) Joback, K. G.; Stephanopoulos, G. Designing Molecules Possessing Desired Physical Property Values. *Proc. FOCAPD '89*; Snowmass, CO, 1989; pp 363−387.

(7) Kier, L. B.; Lowell, H. H.; Frazer, J. F. Design of Molecules from Quantitative Structure−Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts. *J. Chem. Inf. Comput. Sci.* **1993**, *33* (1), 143−147.

(8) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33* (4), 630−634.

(9) Camarda, K. V.; Maranas, C. D. Optimization in Polymer Design Using Connectivity Indices. *Ind. Eng. Chem. Res.* **1999**, *38*, 1884−1892.

(10) Klein, J. A.; Wu, D. T. Computer-Aided Mixture Design with Specified Property Constraints. European Symposium on Computer-Aided Process Engineering-ESCAPE-1. Elsinore, Denmark, 1992, 229−236.

(11) Macchietto, S.; Odele, O.; Omatsone, O. Design of Optimal Solvents for Liquid−Liquid Extraction and Gas Absorption Processes. *Trans. IChemE* **1990**, *69* (A), 429−433.

(12) Maranas, C. D. Optimal Computer-Aided Molecular Design: A Polymer Design Case Study. *Ind. Eng. Chem. Res.* **1996**, *35*, 3403−3414.

(13) Maranas, C. D. Optimization accounting for property uncertainty in polymer design. *Comput. Chem. Eng.* **1997**, *21* (Suppl.), 1019−1024.

(14) Raman, V. S.; Maranas, C. D. Optimization in Product Design with Properties Correlated with Topological Indices. *Comput. Chem. Eng.* **1998**, *22* (6), 747−763.

(15) Vaidyanathan, R.; El-Halwagi, M. M. Computer-aided design of high performance polymers. *J. Elastomers Plast.* **1994**, *26* (3), 277−293.

(16) Vaidyanathan, R.; Gowayed, Y.; El-Halwagi, M. M. Computer-aided design of fiber reinforced polymer composite products. *Comput. Chem. Eng.* **1998**, *22* (6), 801−808.

(17) Vaidyanathan, R.; El-Halwagi, M. M. Computer-Aided Synthesis of Polymers and Blends with Target Properties. *Ind. Eng. Chem. Res.* **1996**, *35* (2), 627−634.

(18) Churi, N.; Achenie, L. E. K. A Novel Mathematical Programming Model for Computer Aided Molecular Design. *Ind. Eng. Chem. Res.* **1996**, *35* (10), 3788−3794.

(19) Ostrovsky, G. M.; Achenie, L. E. K.; Sinha, M. A Reduced Dimension Branch-and-Bound Algorithm for Molecular Design. *Comput. Chem. Eng.* **2003**, *27* (4), 551−567.

(20) Sahinidis, N. V.; Tawarmalani, M.; Yu, M. Design of Alternative Refrigerants via Global Optimization. *AIChE J.* **2003**, *49* (7), 1761−1774.

(21) Lin, B.; Chavali, S.; Camarda, K.; Miller, D. C. Computer-aided molecular design using Tabu search. *Comput. Chem. Eng.* **2005**, *29* 337−347.

(22) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1079−1087.

(23) Douguet, D.; Thoreau, E.; Grassy, G. A Genetic Algorithm for the Automated Generation of Small Organic Molecules: Drug Design using an Evolutionary Algorithm. *J. Comput.-Aided Mol. Des.* **2000**, *14* (5), 449−466.

(24) Kvasnicka, V.; Pospichal, J. Simulated Annealing Construction of Molecular Graphs with Required Properties. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (3), 516−526.

(25) Sundaram, A.; Venkatasubramanian, V. Parametric Sensitivity and Search-Space Characterization Studies of Genetic Algorithms for Computer-Aided Polymer Design. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 1177−1191.

(26) Marcoulaki, E. C.; Kokossis, A. C. Molecular Design Synthesis Using Stochastic Optimization as a Tool for Scoping and Screening. *Comput. Chem. Eng.* **1998**, *22*, S11−18.

(27) Wang, Y.; Achenie, L. E. K. A Hybrid Global Optimization Approach for Solvent Design. *Comput. Chem. Eng.* **2002**, *26*, 1415−1425.

(28) Churchwell, C. J.; Rintoul, M. D.; Martin, S.; Visco, D. P., Jr.; Kotu, A.; Larson, R. S.; Sillerud, L. O.; Brown, D. C.; Faulon, J. L. The signature molecular descriptor. 3. Inverse-quantitative structure−activity relationship of ICAM-1 inhibitory peptides. *J. Mol. Graphics Modell.* **2004**, *22* (4), 263−73.

(29) Faulon, J. L.; Visco, D. P., Jr.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 707−20.

(30) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 1−12.
(31) *Matlab 7, 7.0.4*; MathWorks: 2005.
(32) Bicerno, J. *Prediction of Polymer Properties*. 3rd ed.; Marcel Dekker: New York, 2002.
(33) Contejean, E.; Devie, H. An Efficient Incremental Algorithm for Solving Systems of Linear Diophantine Equations. *J. Inf. Comput.* **1994**, *113* (1), 143−172.
(34) Clausen, M.; Fortenbacher, A. Efficient Solution of Linear Diophantine Equations. *J. Symb. Comput.* **1989**, *8* (1), 201−216.
(35) Faulon, J. L. On using graph-equivalent classes for the structure elucudation of large molecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 338−348.
(36) Faulon, J. L.; Churchwell, C. J.; Visco, D. P., Jr. The signature molecular descriptor. 2. Enumerating molecules from their extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 721−34.
(37) Colbourn, C. J.; Read, R. C. Orderly Algorithms for Generating Restricted Classes of Graphs. *J. Graph Theory* **1979**, *3*, 187−195.
(38) Balaban, A. Chemical Graphs: Looking Back and Glimpsing Ahead. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 339−350.
(39) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45* (4), 839−849.
(40) Guha, R.; Jurs, P. C. Determining the validity of a QSAR model−a classification approach. *J. Chem. Inf. Model.* **2005**, *45* (1), 65−73.
(41) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1912−28.
(42) Brown, W. M.; Martin, S.; Chabarek, J. P.; Strauss, C.; Faulon, J. L. Prediction of β-Strand Packing Interactions using the Signature Product. *J. Mol. Model.* **2005**, in press.

CI0504521