4D-QSPR Analysis and Virtual Screening of Calcite Growth Inhibitor Libraries

José S. Duca and A. J. Hopfinger*

Laboratory of Molecular Modeling and Design (M/C-781), University of Illinois at Chicago, College of Pharmacy, 833 South Wood Street, Chicago, Illinois 60612-7231

Received May 16, 2000. Revised Manuscript Received August 23, 2000

A training set of 35 compounds whose calcite growth inhibition potencies were measured was used to construct a 4D-QSPR model. A site-specific binding pattern of atom types in space, that is a pharmacophore, consisting of six interaction sites between the inhibitors and the surface to which they bind was identified and represented by the 4D-QSPR model. Three of these pharmacophore sites dominate the 4D-QSPR model. One pharmacophore site indicates that its occupancy by any inhibitor atom decreases inhibition potency, suggesting this region of space is occupied by the binding surface. A second pharmacophore site involves an oxygen of a PO_3H_2 group, which is common to all compounds of the training set, hydrogen bonding to the surface. The third major pharmacophore site identifies a nonpolar region of space as being favorable to inhibition potency. A virtual library of 20 analogues to the training set was evaluated by using the 4D-QSPR model as a virtual high throughput screen, VHTS. Seven of the compounds in the virtual library are predicted to be better calcite growth inhibitors than the most potent inhibitor of the training set.

Introduction

Virtual high throughput screening (VHTS) is becoming an increasingly important, and useful, tool to evaluate *virtual compound libraries* (VCL) in the search for both new lead compounds and their structure−activity optimization as part of the preclinical drug discovery process.[1−6] Tens to thousands of compounds can be readily evaluated in VHTS without any of the compounds actually being made or tested. Clearly, there is an enormous time and cost savings to using VHTS in drug-candidate screening, and applications in materials science could be equally advantageous to those being realized in the pharmaceutical sciences.

However, there seems to be very minimal use of VHTS in materials science.

Experimental HTS are nearly nonexistent in materials science largely because of the difficulty in making rapid measurements of the physicochemical properties of interest. There is no analogue measurement in materials science applications to the rapid binding assay that is common in pharmaceutical HTS. On the other hand, the construction of a VHTS in materials science has the same requirements as those of pharmaceutical applications: (a) a training set of compounds and corresponding measures of the property of interest, (b) a pool of trial descriptors for the compounds of the training set, and (c) a means of constructing a statistical relationship, or correlation, between the property measures and some subset of the trial descriptor pool.

The resulting relationship between the property of interest and the subset of descriptors is usually referred to as a quantitative structure−property relationship, or QSPR. The application of the QSPR to predict the values of the property of interest for a library of hypothetical compounds transforms the QSPR into a VHTS. In other words, a QSPR model can be used as a VHTS.

Unfortunately, not all QSPR models of equal statistical significance for a training set are actually "equal". That is, the predictive power of equal QSPR models *outside* the training set can be different. The utility of a QSPR model as a VHTS will depend on its predictive power. Normally, the closer the descriptors of a QSPR model are to reflecting the actual mechanism of action, the more predictive power the model will possess. Thus, there is an impetus to understand, or simply postulate, a mechanism of action, and to generate corresponding descriptors as part of a QSPR analysis. If the resulting QSPR model has poor predictive power, it and the mechanism of action can be abandoned and an alternate mechanism, and corresponding set of descriptors, used in another QSPR analysis.

A general mechanism encountered in materials science applications is the binding of small molecules to

* Corresponding author. Phone: 312-996-4816. Fax: 312-413-3479. E-mail: hopfingr@uic.edu.

(1) 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.

(2) Hopfinger, A. J.; Duca, J. S. Extraction of pharmacophore information from high-throughput screens. *Curr. Opin. Biotechnol.* **2000**, *11*, 97−103.

(3) Murray, C. M.; Cato, S. J. Design of libraries to explore receptor sites. *J. Chem. Inf. Computer Sci.* **1999**, *39*, 46−50.

(4) Pickett, S. D.; Luttmann, C.; Guerin, V.; Laoui, A.; James, E.; DIVSEL and COMPLIB−strategies for the design and comparison of combinatorial libraries using pharmacophoric descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 144−150.

(5) Bures, M. G.; Martin, Y. C.; Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.

(6) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design and compound selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.

10.1021/cm000398y CCC: $19.00
Published on Web 11/22/2000

surfaces. Virtually all coatings, paint, and printing applications fall into this mechanistic classification. These binding processes are very similar in modeling format to the ligand−receptor binding interaction that is fundamental to pharmaceutical science. Thus, the quantitative structure−activity relationship, QSAR, methods used to model ligand−receptor binding should find use in materials science small molecule−surface binding applications. In the study reported here the inhibition of calcite crystal growth by small organics has been modeled by 4D-QSAR analysis,[7] which is a methodology that has been used to develop QSAR models for ligand−receptor interactions.[8−10] The resultant 4D-QSPR models for inhibition of calcite growth were used to generate the "active" conformation of each calcite inhibitor in the training set and as VHTS to evaluate a small library of hypothetical calcite growth inhibitors.

Method

A. Training Set of Calcite Crystal Growth Inhibitors. A training set of 35 calcite crystal growth inhibitors, along with their respective measured calcite crystal growth inhibitor "constants", I_g, were used to construct the 4D-QSPR models. The I_g values are actually crystal growth inhibition times (in minutes) determined from a supersaturated solution of calcium and carbonate at constant temperature.[11] Solution mixing is done to ensure metastability at a desired ionic strength and pH (≈8.5). Calcium seed crystals of a known fixed size (≈10 μm per side) are added to the solution. As the experiment runs, calcium and carbonate/bicarbonate titrants are fed to maintain the fixed pH. Inhibitor is added after a small amount of titrants have been fed so that the initial growth rate can be observed prior to the addition of inhibitor. The consumption of titrants is measured as a function of time to obtain the growth rate. The inhibition time, I_g, is taken as the time required for the repropagation of crystal growth after the inhibitor has been added to the system. Multiple experiments are normally performed, and corrections (normalization) are made for variations in initial crystal growth rates and sizes of the seed crystals.

The structures of the compounds and their I_g values [$-\log(I_g)$ measures are used in constructing the 4D-QSAR models] are given in Table 1. The compounds all contain at least one $-PO_3H_2$ group. Thus, to some degree the compounds can be considered analogues to one another. On the other hand, the compounds are markedly varied in size and composition of atoms so that they possess considerable structural diversity. Overall, the training set can be considered to cover a relatively wide range of chemical structure within the family of compounds having $-PO_3H_2$ groups.

4D-QSPR models were constructed using $-\log(I_g)$ as a dependent variable measure which monitors the free energy change associated with calcite crystal growth inhibition.[12]

B. The 4D-QSAR/QSPR Formalism. The 4D-QSAR/QSPR formalism and corresponding methodology has been presented in detail.[7−10] A commercial software package to perform 4D-QSAR and 4D-QSPR analyses is available, and its operations manual[13] is also a good reference that describes the methodology.

In summary, each compound of the training set from which the 4D-QSPR models are constructed is sampled with respect to its conformational freedom. Molecular dynamics simulation, MDS, is currently used to generate the conformational ensemble profile of each molecule.[7,14] In this study 100 000 conformations were sampled in the MDS. Each conformation of the ensemble profile of each molecule is then placed in a grid cell space of some particular resolution according to some selected alignment. By way of an example, compound **10** of Table 1 is shown in Figure 1 in its grid cell space for a particular three-ordered atom alignment. That is, these common atoms are selected to perform a quantitative spatial comparison of molecules, and these atoms are selected in a particular sequence in forming the comparison rule.

The grid cell occupancies of the set of atom types, defined in Table 2, composing each molecule of the training set are recorded for each conformation of the molecule in its ensemble profile. The resulting composite set of grid cell occupancy values for each of the atom types composing the molecule, called *grid cell occupancy descriptors, GCODs*, become the pool of independent variables for constructing 4D-QSPR models. Non-GCOD descriptors can also be selected by the user and added to the GCOD descriptor pool. No non-GCOD descriptor was found to be significant in this study and, consequently, this class of descriptors is not discussed any further.

The number of GCODs in the descriptor pool can be very large in comparison to the size of the training set. Hence, a data reduction is done as part of the 4D-QSPR model construction process. Partial least squares, PLS, regression[15] and some filtering rules[7−9] are applied to the complete set of descriptors in the pool to determine a set of only the most highly weighted (significant) GCOD descriptors for consideration in further 4D-QSPR model building. Currently, the 200 most highly

(7) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(8) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; deAlencastro, R. B. Four-dimensional quantitative structure−activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−938.

(9) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand−receptor binding free energy by 4D-QSAR analysis: application to a set of glucose inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141−1150.

(10) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase *b*. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160.

(11) Reed, P. E., Kamrath, M. A., Carter, P. W., Davis, R. Y. Ether diphosphonate scale inhibitors. U.S. Patent 5,772,893, June 30, 1998.

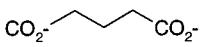
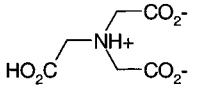
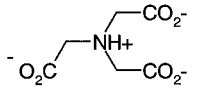
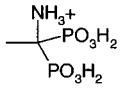
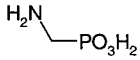
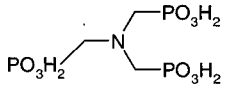
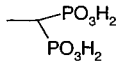
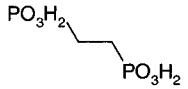
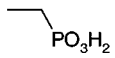
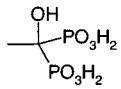
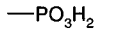
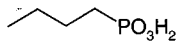
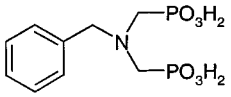
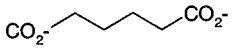
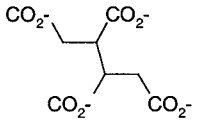
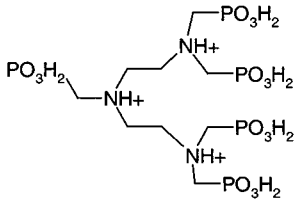
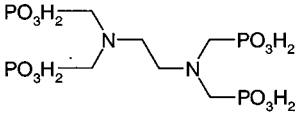
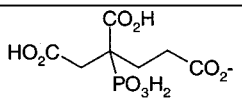
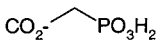
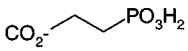
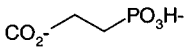
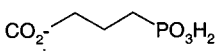
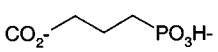
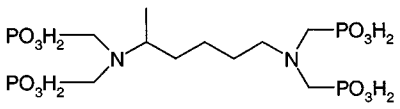
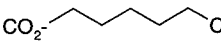
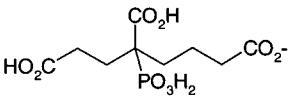
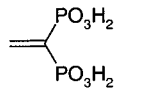
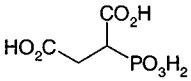
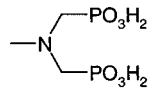
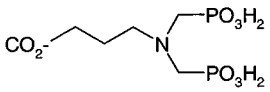
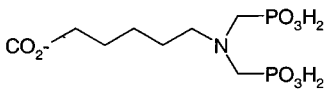
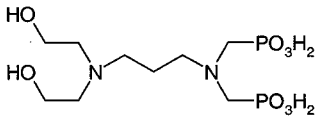
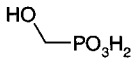
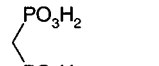
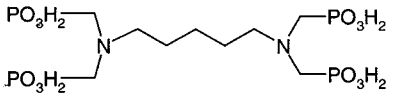
(12) Tokarski, J. S.; Hopfinger, A. J. Prediction of ligand−receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis. Application to a set of peptidometic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792−811.

(13) 4D-QSAR User's Manual, Version 2.0; The Chem21 Group, Inc.: 1780 Wilson Drive, Lake Forest, IL 60045; 2000.

(14) Doherty, D. C. MOLSIM User's Guide; The Chem21 Group, Inc.: 1780 Wilson Drive, Lake Forest, IL 60045; 1998.

(15) Glen, W. G.; Dunn, W. J., III; Scott, D. R. Principal components analysis and partial least squares. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−354.

Table 1. Chemical Structures and Corresponding Calcite-Growth Inhibitor Measures, −log(I_g), Used in the 4D-QSPR Analysis

Comp	Structure	-log(I_g)
1.		0
2.		0
3.		0
4.		2.845
5.		0
6.		1.699
7.		3.079
8.		0
9.		0
10.		3.114
11.		0
12.		0
13.		0
14.		0
15.		0
16.		2.511
17.		2.602
18.		2.531
19.		
20.		1.845
21.		0
22.		0
23.		0
24.		2.176
25.		0
26.		2.531
27.		2.875
28.		2.875
29.		0
30.		0
31.		0
32.		0
33.		1.875
34.		3.041
35.		2.550

weighted GCODs are used as an input basis set for 4D-QSPR model optimization using a particular genetic algorithm called the "genetic function approximation", GFA.[16,17]

Assigning grid cell occupancy, that is, determining a set of GCODs, for the ensemble set of conformations of each molecule in the training set can be repeated for another alignment. This will, in turn, permit the construction of an optimum 4D-QSPR model for the new alignment. In other words, the single ensemble set of conformations of each molecule in the training set can be used to generate optimum 4D-QSPR models for each

(16) Rogers, D. G/SPLINES: A hybrid of Friedman's multivariate adaptive regression splines (MARS) algorithm with Holland's genetic algorithm. *The Proceedings of the Fourth International Conference on Genetic Algorithms*; San Diego, 1991; pp 38−46.

(17) Rogers, D.; Hopfinger, A. J. Applications of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

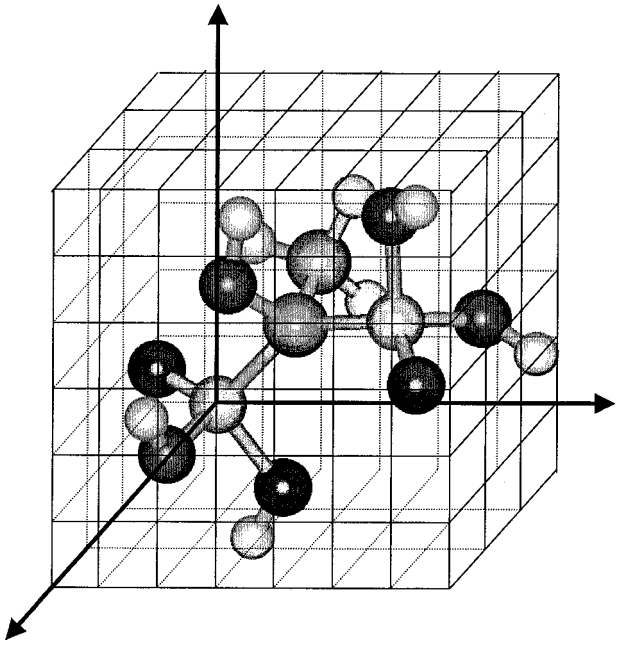


Figure 1. Compound **10** of Table 1, the most potent calcite growth inhibitor of the training set, shown in a 1A grid cell lattice space.

Table 2. Grid Cell Occupancy Atom Types and Corresponding Numerical Coding

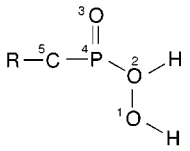
grid cell descriptor code	atom type
0	all
1	nonpolar
2	polar positive
3	polar negative
4	H-bond acceptor
5	H-bond donor
6	aromatic

alignment selected, and an arbitrarily large number of alignments can be selected. Thus, 4D-QSPR model optimization can be performed as a function of alignment, GCODs, and/or conformational sampling.

It is possible to hypothesize an "active conformation" of each compound in the training set. This is achieved by first identifying all conformer states in the sampling of a molecule that are within ΔE (currently set at 2 kcal/mol) energy units of the global minimum of the conformational ensemble profile. Each member of the resulting set of low-energy conformations is individually evaluated by using the best 4D-QSPR model. Since only a ***single*** conformation is used, the grid cell occupancy is either zero or one for each GCOD of the model. The single conformation within ΔE that predicts the highest "activity" (calcite crystal growth inhibition constant, I_g) is selected as the active conformation. The postulated active conformations can be used as structure design templates.

The difference between the predicted activity (using the active conformation) and the observed activity of a molecule can be viewed as the loss in possible activity of the molecule due to its flexibility. That is, the active conformation tends to have appropriate atom types that completely occupy activity-enhancing GCODs, while the activity decreasing GCODs are not occupied. In contrast, the conformational ensemble profile of the molecule leads to both the activity enhancing and activity decreasing GCODs being partially occupied. Said another way, the difference between predicted activity, using the active conformation, and the observed activity is the loss in activity due to conformational entropy.

Table 3. Summary of the Statistical Fitting Measures for the 10 Best Models Obtained from the 4D-QSPR Analysis



alignment 1[a] 3 4 1
alignment 2[a] 2 3 5

	alignment 1				alignment 2			
model	LOF	LSE	R^2	Q^2	LOF	LSE	R^2	Q^2
1	1.07	0.25	0.85	0.81	1.34	0.44	0.74	0.67
2	1.17	0.28	0.83	0.78	1.93	0.46	0.73	0.63
3	0.89	0.29	0.83	0.78	1.16	0.50	0.70	0.63
4	1.24	0.29	0.83	0.78	1.53	0.50	0.70	0.63
5	0.91	0.30	0.82	0.78	1.21	0.52	0.69	0.62
6	0.96	0.23	0.86	0.77	1.61	0.53	0.69	0.62
7	0.92	0.30	0.82	0.77	1.83	0.43	0.74	0.62
8	1.11	0.26	0.84	0.76	1.50	0.49	0.71	0.62
9	0.93	0.31	0.82	0.76	1.25	0.54	0.68	0.61
10	0.99	0.32	0.81	0.76	2.69	0.43	0.74	0.61

[a] In case of molecules **1**, **2**, **3**, **14**, **15**, and **25**, both alignments refer to the oxygen atoms of carboxylate groups.

The 4D-QSPR model can also be used as a *virtual high throughput screen*, VHTS, to predict the activities of the members of a virtual library of "similar" compounds to those of the training set.[18] There is no rigorous way to define what "similar" means in terms of accurately being able to estimate activity of a compound outside of the training set. A safe course of action is to simply expand the analogue characteristics of the training set. In this application a small virtual library was generated where the library members are analogues to the smaller, but also the better, inhibitors of the training set.

The use of a 4D-QSPR model as a VHTS is not novel, but rather an extended application of a fundamental purpose of constructing the QSPR model. QSPR models are constructed to permit forecasting the measure of a property of interest for hypothetical molecules of interest.

Results

Two alignments, defined in Table 3 using a general compound able to encompass most of the compounds present in Table 1, were evaluated in the 4D-QSPR analysis. One reason only two alignments were considered is that a very good model was constructed for one of the alignments. The other reason for limiting alignment choices is although there is a relatively wide range in size (number of atoms) of the molecules in the training set, only a small grouping of atoms is common across the compounds of the training set. It was decided to limit the alignments to three-ordered atom alignments within this common grouping. Table 3 also contains, among other statistical measurements, the

(18) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand–receptor binding free energy by 4D-QSAR analysis: application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 9, 1141–1150.

Table 4. Ten Best 4D-QSPR Models for Alignment 1 for the Calcite Growth Inhibition, $-\log(I_g)$, Training Set

1	$-\log(I_g) =$ $-$ 24.55 * GC1(a) $+$ 4.37 * GC2(hbd) $+$ 3.20 * GC3(a) $+$ 24.71 * GC4(p$-$) $+$ 26.80 * GC5(hbd) $+$ 10.95 * GC6(hba) $-$ 0.065	**6**	$-\log(I_g) =$ $-$ 11.34 * GC12(a) $+$ 7.77 * GC2(hbd) $+$ 11.15 * GC13(p+) $+$ 24.15 * GC4(p$-$) $+$ 27.05 * GC5(hbd) $+$ 12.18 * GC6(hba) $-$ 0.084
2	$-\log(I_g) =$ $-$ 16.36 * GC7(np) $+$ 4.60 * GC2(hbd) $+$ 2.98 * GC3(a) $+$ 24.49 * GC4(p$-$) $+$ 25.74 * GC5(hbd) $+$ 10.99 * GC6(hba) $-$ 0.058	**7**	$-\log(I_g) =$ $-$ 11.92 * GC12(a) $+$ 7.70 * GC2(hbd) $+$ 24.66 * GC4(p$-$) $+$ 27.04 * GC5(hbd) $+$ 13.56 * GC6(hba) $-$ 0.084
3	$-\log(I_g) =$ $-$ 31.80 * GC1(a) $+$ 6.50 * GC3(a) $+$ 20.09 * GC8(a) $+$ 26.65 * GC5(hbd) $+$ 7.82 * GC6(hba) $-$ 0.046	**8**	$-\log(I_g) =$ $-$ 31.36 * GC1(a) $+$ 8.44 * GC14(hba) $+$ 6.49 * GC3(a) $+$ 20.13 * GC8(a) $+$ 26.98 * GC5(hbd) $+$ 7.62 * GC6(hba) $-$ 0.091
4	$-\log(I_g) =$ $-$ 9.56 * GC9(a) $+$ 5.30 * GC2(hbd) $+$ 2.24 * GC3(a) $+$ 9.83 * GC10(p$-$) $+$ 26.66 * GC5(hbd) $+$ 11.50 * GC6(hba) $-$ 0.037	**9**	$-\log(I_g) =$ $-$ 14.98 * GC15(a) $+$ 7.60 * GC2(hbd) $+$ 22.74 * GC4(p$-$) $+$ 25.39 * GC5(hbd) $+$ 13.57 * GC6(hba) $-$ 0.027
5	$-\log(I_g) =$ $-$ 35.97 * GC11(np) $+$ 6.45 * GC3(a) $+$ 17.46 * GC8(a) $+$ 25.81 * GC5(hbd) $+$ 7.64 * GC6(hba) $-$ 0.026	**10**	$-\log(I_g) =$ $-$ 22.68 * GC16(np) $+$ 6.29 * GC3(a) $+$ 23.47 * GC4(p$-$) $+$ 25.42 * GC5(hbd) $+$ 8.05 * GC6(hba) $-$ 0.005

cross-validated correlation coefficient, Q^2, of the top-10 4D-QSPR models found for each alignment. It is clear that alignment 1 of Table 3 is superior, as measured by the Q^2 of its 10 best 4D-QSPR models, to the other alignment. Hence, only alignment 1 was considered in the further analysis of the training set and in performing a VHTS.

The top-10 4D-QSPR models of alignment 1 are summarized in Table 4. These models were obtained by using the GFA. The best 4D-QSPR model is

$$-\log(I_g) = -24.6\mathrm{GC1(a)} + 4.37\mathrm{GC2(hba)} + 3.20\mathrm{GC3(a)} + 24.7\mathrm{GC4(np)} + 26.8\mathrm{GC5(hbd)} + 10.9\mathrm{GC6(hbd)} - 0.065 \quad N = 35 \quad R^2 = 0.85 \quad Q^2 = 0.81 \quad (1)$$

In eq 1 GCI(x) is the Ith significant GCOD descriptor having the x-type IPE as defined and coded in Table 2. Figure 2 is a plot of observed versus predicted $-\log(I_g)$ values for the compounds of the training set. No compound in the training set is an outlier.

To determine if the top-10 4D-QSPR models of alignment 1 are providing common, or distinct, structure-inhibition information, the correlation coefficients of the residuals in fit of the 4D-QSPR models were computed and are reported in Table 5. Highly similar models will have highly similar residuals of fit to the observed dependent variables of the training set. Correspondingly, these residuals of fit will be highly correlated to one another. All of the top-10 4D-QSPR models have residuals of fit which are highly correlated to one another ($R > 0.85$). Therefore, there is only **one** 4D-QSPR model, namely, eq 1, which is the best 4D-QSPR model of the top-10, to explain the training set data.

A linear cross-correlation matrix of the GCODs of eq 1 as well as $-\log(I_g)$ has been built and is given in Table 6. No pair of GCODs of eq 1 are more highly correlated than an $R = 0.57$. This finding suggests that each of the GCODs is largely independent of the others and is providing distinct information to explain the data comprising the training set. GCODs GC3(a) and GC5(hbd) are, individually, the most highly correlated descriptors to $-\log(I_g)$.

Figure 3 shows another way to analyze the presence of important GCODs in the optimum model. The frequency of occurrence of each GCOD in the top-10 models is shown in Figure 3, and it can be seen that only GC5(hbd) is present in all 10 models, but also that five out of six GCODs present in model 1 are also present in 60% of the other nine models.

The predicted active conformation of each compound of the training set was determined from its conformational ensemble profile and eq 1 according to the scheme given in the Method section. The GCODs of the 4D-QSPR model (eq 1) are shown mapped into space relative to the postulated active conformation of a good inhibitor, compound 10 of Table 1, in Figure 4. The grid cells are represented by spheres whose diameters are equal to the grid cell resolution, namely, 1 Å. It can be seen in Figure 4, for example, that GC5(hbd), one of the most important GCODs in the 4D-QSPR model, as judged by its regression coefficient, its individual correlation to $-\log(I_g)$, and its frequency of appearance in all top-10 4D-QSPR models, is located near the protonated oxygens of one of the two PO_3H_2 groups. The nearest proton is located 1.24 Å from GC5(hbd). GC1(a), the other highly significant GCOD of eq 1, is not near any atom of compound **10**. Occupancy of this grid cell by any type of atom is predicted to be detrimental to good inhibition potency since the sign of the regression coefficient of GC1(a) is negative in eq 1.

GC6(hba) is another important GCOD in the best 4D-QSPR model because of its large regression coefficient. This GCOD is only 1.04 Å from the nonprotonated oxygen of one of the PO_3H_2 moieties. The other two GCODs of eq 1 relative to compound 10 are GC2(hbd), located only 0.59 Å from a protonated oxygen, and GC3(a), which is 1.34 Å from the methyl group.

Figure 5 is the same as Figure 4 except that the reference molecule, in its active conformation, is compound **13** of Table 1, which is a poor calcite growth inhibitor. In this case GC5(hbd) is far from any hydrogen bond donor group, but 2.8 Å from a carbon atom. GC6(hba) is the only GCOD of eq 1 that is located near a "correct" atom type, being 1.4 Å from an oxygen atom. GC2(hbd) is about 2.4 Å from an oxygen of compound **13**.

These types of visual comparisons between compounds in the training set, in their predicted active conformations, and the GCODs of the corresponding 4D-QSPR model is a qualitative approach to validating the model. Moreover, the graphical display of the GCODs overlaid on a good inhibitor, as in Figure 4, can be used

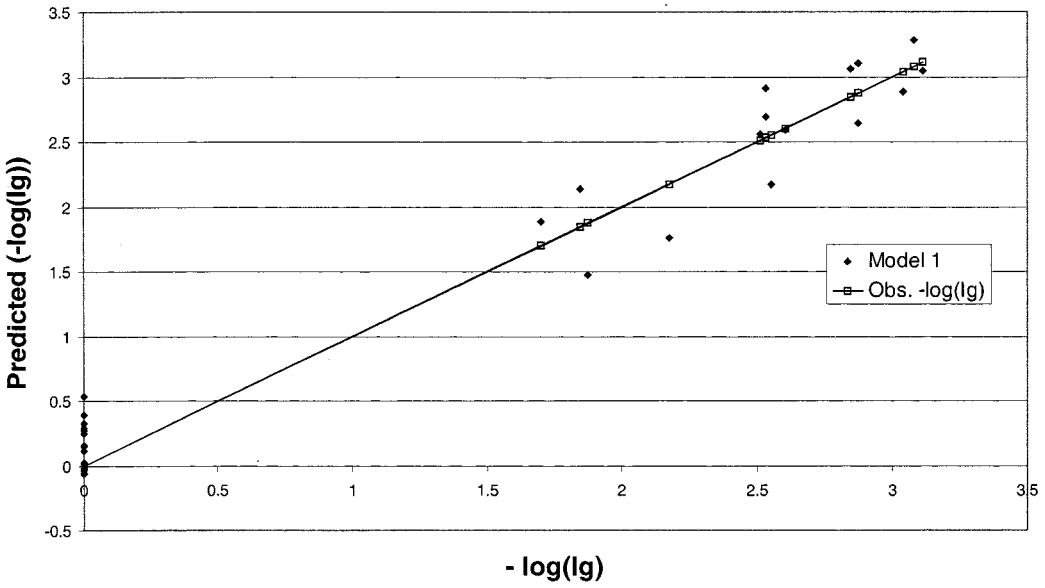


Figure 2. Plot of the predicted values for [$-\log(I_g)$] versus the corresponding experimental values. The ideal relationship is displayed with a trend line of unitary slope.

Table 5. Cross-Correlation Matrix for the Residuals (Res) of Error in Predicting Activity Data of the Best 10 Models for Alignment 1[a]

	Res1	Res2	Res3	Res4	Res5	Res6	Res7	Res8	Res9	Res10
Res1	1.00									
Res2	0.98	1.00								
Res3	0.93	0.90	1.00							
Res4	0.98	0.97	0.87	1.00						
Res5	0.92	0.90	0.99	0.87	1.00					
Res6	0.83	0.84	**0.70**	0.87	**0.68**	1.00				
Res7	0.94	0.94	**0.77**	0.96	**0.75**	0.87	1.00			
Res8	0.84	0.83	0.95	**0.79**	0.94	**0.75**	**0.67**	1.00		
Res9	0.91	0.94	**0.74**	0.95	**0.74**	0.84	0.96	**0.64**	1.00	
Res10	0.92	0.93	0.98	0.88	0.98	**0.72**	**0.77**	0.94	**0.78**	1.00

[a] All pairs with $R < 0.80$ are presented in bold.

as a qualitative design template to see how well a hypothetical molecule fits the spatial binding pattern of atom types defined by the GCODs of the 4D-QSPR model.

Prediction of the $-\log(I_g)$ value of a member of the training set using a 4D-QSPR model, but only employing the *predicted* active conformation, permits an estimate of the conformational entropy contribution to inhibition potency of the compound. Basically, the difference between the predicted and the observed inhibition potency is due to conformational entropy. The differences between predicted and observed inhibition potencies are plotted in Figure 6 for all the compounds of the training set. The compound numbering is the same as given in Table 1. Caution needs to be exercised in interpreting these differences as the *quantitative* gains and/or loses in $-\log(I_g)$ arising from conformational entropy. The effect of the occupancy differences between the actual CEP occupancy values of a compound and the limiting occupancy values of 0 or 1 for the postulated active conformation can be attenuated by the regression coefficients which are derived solely on the basis of grid cell occupancy.

In general, most of the differences in Figure 6 are negative, suggesting that there is a loss in possible inhibition potency for most compounds. Only four of the 35 compounds of the training set have positive inhibition potency differences. The conformational flexibility, of which conformational entropy is a measure, seemingly prevents the "binding state" from being fully realized owing to the population of conformational states accessible to the inhibitor.

Compounds **12** and **21** both have extremely large negative differences between observed and predicted $-\log(I_g)$ values. Compounds **12** and **21** both do *not* occupy GC1 (negative regression coefficient in eq 1) for their postulated active conformations, which is the primary source of the very large negative differences in inhibition potency. Compound **4** has a very large positive difference in Figure 6 because it does *not* occupy GC5 of eq 1. The regression coefficient of GC5 is very large so that a fractional occupancy of this GCOD by compound **4** yields a significant contribution to its inhibition potency.

The information in Figure 6 suggests not using compound **4** as a lead compound for generating analogues having its core structure. The inability to occupy GC5 for preferred low-energy conformations and binding alignment limits its inherent inhibition potency. Conversely, compounds **12** and **21** might be considered as

Table 6. Cross-Correlation Matrix of the GCOD and $-\log(I_g)$ Pairs of the Optimum 4D-QSPR Model

	GC1(a)	GC2(hdb)	GC3(a)	GC4(p-)	GC5(hbd)	GC6(hdb)	$-\log(I_g)$
GC1(a)	1.00						
GC2(hdb)	−0.08	1.00					
GC3(a)	0.61	0.57	1.00				
GC4(p−)	0.37	−0.06	0.53	1.00			
GC5(hbd)	−0.03	−0.07	−0.13	−0.07	1.00		
GC6(hdb)	0.07	−0.17	0.25	−0.03	−0.10	1.00	
$-\log(I_g)$	0.17	0.41	0.57	0.23	0.43	0.42	1.00

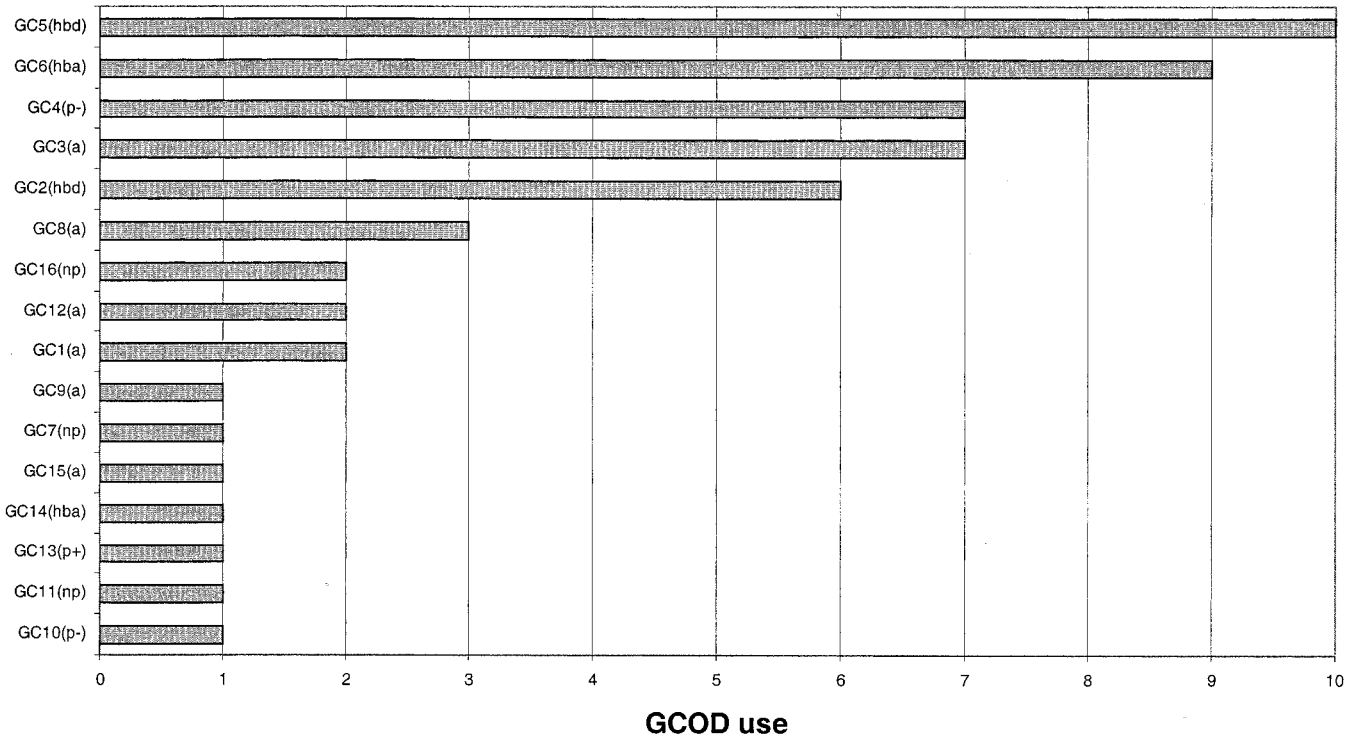


Figure 3. Bar graph representation of the frequency of use for the most relevant GCODs in the top-10 models obtained from the 4D-QSPR analysis.

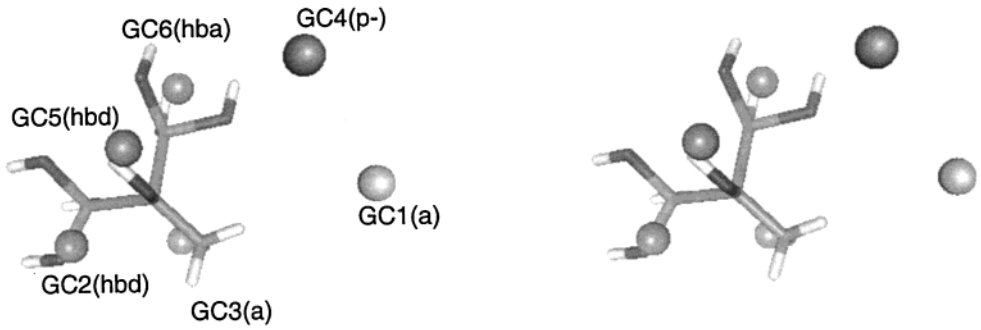


Figure 4. Stereorepresentation of a stick model for compound **10** (the most active calcite growth inhibitor) and the best 4D-QSPR model 1, given by eq 1. The GCODs are represented as "floating" spheres: GCODs which enhance inhibition are darker in shading, while those GCODs which diminish inhibition potency are lighter in shading.

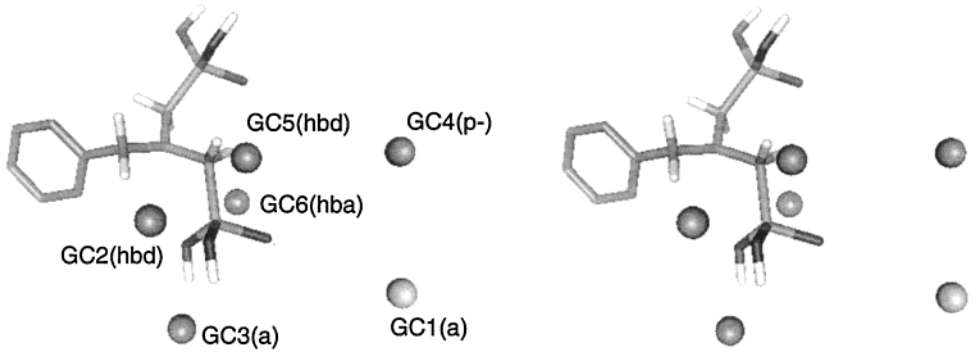


Figure 5. Same as Figure 4, but compound **13**, one of the least potent calcite growth inhibitors, is used as the reference compound.

lead structures, since they both do not occupy GC1. However, they also do not occupy inhibition potency enhancing GCODs and are inactive! Thus, compounds **4**, **12**, and **21**, which are anomalies in terms of the behavior of $-\log(I_g)$ and conformational entropy, are best avoided as lead sources for generating new libraries.

A small virtual library of 20 compounds was constructed on the basis of the high inhibition potencies, and small sizes, of compounds **7** and **10** of Table 1. This virtual library is given in Table 7. Each of the 20 virtual compounds was screened by the 4D-QSPR model, expressed by eq 1, as the VHTS. Virtual screening involves processing each virtual test compound in the same

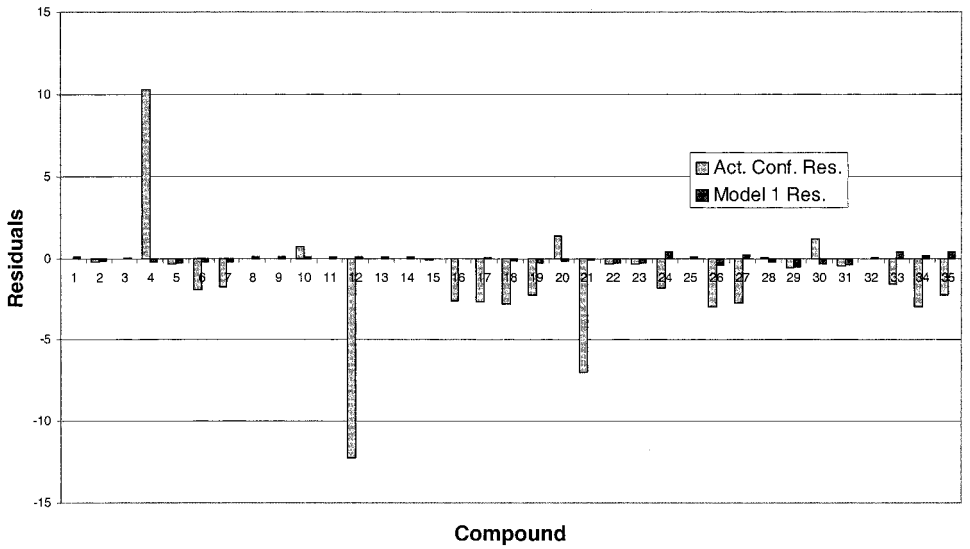
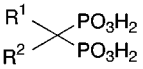


Figure 6. Graphical representation of the residuals in the $-\log(I_g)$ values between those of the predicted active conformation using eq 1 and the corresponding observed values. The residuals obtained using the GCOD values in eq 1 and the observed values are also shown for comparison.

Table 7. Virtual Library of Calcite Growth Inhibitors Based on Compounds 7 and 10 of the Training Set

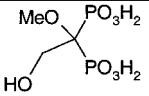
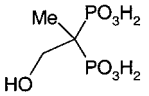
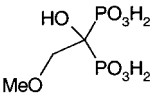
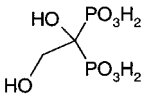
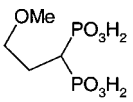
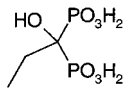
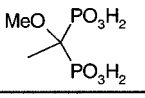


compd	R1	R2
1	OMe	Me
2	Me	Me
3	H	CH_2OH
4	H	Et
5	H	CH_2OMe
6	H	$CH_2NH_3^+$
7	H	OH
8	H	NH_3^+
9	H	OMe
10	CN	Me
11	H	CN
12	H	CH_2CN
13	H	Prop
14	H	CH_2CH_2OH
15	H	CH_2CH_2OMe
16	OH	CH_2OH
17	OH	CH_2OMe
18	OMe	CH_2OH
19	OH	Et
20	Me	CH_2OH

manner as done for each compound in the training set. This processing includes computing the GCOD values of the terms in eq 1 by generating the conformational ensemble of each virtual test compound.

Seven of the virtual compounds were predicted by the VHTS to be more potent calcite growth inhibitors than compound **10** of Table 1, which is the most potent inhibitor in the training set. These seven compounds, and their predicted $-\log(I_g)$ values, are given in Table 8. The high predicted inhibitor potencies of these compounds are largely due to the high occupancy of GC2(hba) by a nonprotonated oxygen of a PO3H2 group, minimal occupancy of GC1(a) by any atom of the inhibitor, and partial occupancy of GC3(a) by either a methyl group or a hydroxyl group.

Table 8. Calcite Crystal Growth Inhibitors from the Virtual Library, Which Are Predicted To Be Better Inhibitors Than the Most Potent Inhibitor, Compound 10, of the Training Set (The 4D-QSPR Model, eq 1, Has Been Used as the VHTS)

Compound No.	Structure	Predicted $-\log(I_g)$
18		3.82
20		3.56
17		3.47
16		3.36
15		3.29
19		3.28
1		3.28

Discussion

The ability to establish a significant 4D-QSPR model devoid of any nonspecific binding property, and only involving GCODs, suggests that calcite crystal growth inhibition does involve some degree of *specific binding* of the inhibitor. Possible molecular mechanisms of crystal growth inhibition include (a) complexing of the inhibitor with a calcium ion and/or calcium carbonate/bicarbonate in solution, (b) deposition of the inhibitors on the surfaces of growing calcite crystals (the result of such inhibitor deposition would create defects in the crystals and limit growth), or (c) a combination of both mechanisms described above.

Deposition of inhibitors on crystal surfaces is more consistent with specific binding, as is implied by the 4D-

QSPR model found in this study, than inhibitor complexing to free calcium ions in solution. There appears to be multiple ways to complex free calcium ions to the inhibitors of the training set. Further, no single geometric mode of free calcium ion complexing to inhibitors can be identified which separates good inhibitors from poor inhibitors. Still, no arrangement of calcium ions on the common growth faces of calcite match the pattern of GCODs in the 4D-QSPR model. Perhaps the highly ordered packing of a crystal structure imparts large steric restrictions to the way in which an inhibitor can fit, and bind, to sites on a *growing* surface of a calcite crystal. The 4D-QSPR model captures the representative average requirements of an inhibitor to bind to these sites on the growing surfaces and inhibit growth.

An inspection of the best 4D-QSPR model in both its numerical (eq 1) and graphic (Figures 4 and 5) representations delineates the pattern of atoms of the inhibitors (normally called a pharmacophore in biological applications) primarily responsible for inhibition potency, or lack thereof. The pharmacophore for calcite growth inhibition derived from the training set identifies six possible interaction sites on an inhibitor which can influence its potency. Of these six sites, a steric restriction, probably involving some group on the binding surface, is associated with GC1(a). A favorable nonpolar binding site on the surface is inferred from GC4(np), and one of the protonated oxygens of one of the PO_3H_2 groups is predicted to hydrogen bond to a surface hydrogen bond acceptor on the basis of GC5(hbd). These three pharmacophore sites dominate (largest regression coefficients in eq 1) inhibition specificity in the 4D-QSPR model. At least two other GCOD descriptors in eq 1 are important to predict inhibition potencies of compounds of the virtual library: GC2(hba) for hydrogen bonding of a nonprotonated oxygen of a PO_3H_2 moiety to a proton on the binding surface and GC3(a) that describes the benefit of a group of comparable size to a methyl or hydroxy group to enhancing inhibition potency.

The seven virtual compounds of the VHTS predicted to be potent inhibitiors, which are given in Table 8, suggest that many of the compounds in the training set are needlessly large and "complex" in chemical structure relative to the requirements for realizing high inhibition potency. The compounds in Table 7 also illustrate that a wide variety of substituent combinations can be placed on the core of compound **10** of Table 1 in order to explode the exploration of inhibition potency in this class of simple molecules. Moreover, exploration of this class of compounds appears to be a more fruitful endeavor, given the findings in Table 8, than expanding the chemistry away from direct analogues as is characteristic of the training set.

This calcite crystal growth inhibition application study suggests that 4D-QSPR analysis, and the subsequent use of a 4D-QSPR model as a VHTS, may comprise a useful tool to develop new materials which act through *site specific* interaction mechanisms. The utility of applying 4D-QSPR analysis to other classes of material design problems, not explicitly involving geometric mechanistic specificity, is not known. However, the 4D-QSPR paradigm does allow a useful way of cataloging and processing information about molecular geometry as a function of conformational freedom, time, temperature, and molecular alignment. Any problem in materials design that can be formulated in terms of these degrees of freedom could, presumably, benefit from 4D-QSPR analysis.

Acknowledgment. We appreciate both the financial and technical support of the Nalco Chemical Company. Resources of The Laboratory of Molecular Modeling and Design at UIC were employed in performing this research. We appreciate the helpful and stimulating discussions with Drs. D. Johnson, P. Young, and V. Narutis of Nalco over the course of this study.

CM000398Y