Construction of a Virtual High Throughput Screen by 4D-QSAR Analysis: Application to a Combinatorial Library of Glucose Inhibitors of Glycogen Phosphorylase *b*

Anton J. Hopfinger,* Andrea Reaka, Prabha Venkatarangan,† José S. Duca, and Shen Wang

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

Received April 5, 1999

The 4D-QSAR model developed for a training set of 47 glucose analogue inhibitors of glycogen phosphorylase, and reported in the previous paper in this issue, was used as a basis for developing virtual high throughput screen, VHTS, models to screen a focused combinatorial virtual library of 225 additional inhibitors. Techniques to develop, evaluate, and apply VHTS models derived from 4D-QSAR models are presented. Application of the VHTS models to screen the virtual library results in the prediction of compounds which bind both more, and less, strongly to the enzyme than the best and worst binders of the training set. Analysis of the binding predictions across the virtual library reveals patterns of structure−activity information that can be useful to design new focused libraries. The possible use of overfit QSAR models, with respect to the training data set, as VHTS models is discussed and explored.

INTRODUCTION

Virtual compound libraries and virtual high throughput screens (VHTSs) taken together offer a computational approach that has the potential to prioritize and guide library synthesis in both lead optimization and identification studies. In this paper we describe the use of a new QSAR method, termed 4D-QSAR analysis,[1−3] to construct useful VHTS models for the screening of a virtual library of compounds, particularly analogues. This approach to VHTS requires that a relevant training set of compounds (and corresponding bioactivities), usually a subset of the virtual library to be studied, is available to construct the 4D-QSAR models used as the VHTS.

In preclinical applications seeking drug candidates a limited number of compounds are made and tested. The results are sufficiently encouraging to suggest expanding the compound data set by combinatorial methods. The question is what substituents, "decorations", and/or "core changes" should be incorporated into the new combinatorial chemistry libraries. A QSAR derived from the original structure−activity data set, the training set, can be used as a VHTS to guide the design of target combinatorial chemical libraries. Compounds of a virtual library that are subsequently made and screened for biological activity can be added to the original training set. Thus, the VHTS, based on QSAR analysis, can be evolved to take advantage of all available SAR data. Also, a unique VHTS can be constructed from QSAR analysis for each biological activity end point being investigated.

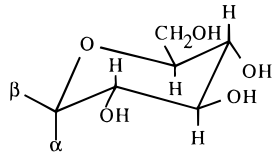
Actually, use of a QSAR model as a VHTS is no different from the "traditional" application of QSAR analysis. The goal of constructing a QSAR is to apply it as a guide in further extending the SAR data set toward optimal activity. However, not all QSARs of equal statistical significance are "created equal". In particular, the basis set of descriptors used to construct the QSAR model can dictate, and limit, the utility of the application of the QSAR. The QSAR model descriptor set becomes particularly important when the QSAR is used as a VHTS in compound library design and evaluation. A good in vitro VHTS should have two components: (A) the spatial pharmacophore responsible for the binding of a ligand to the receptor and (B) the spatial sites not available [due to the receptor] to the ligand for receptor binding. A 4D-QSAR model contains both of these features as represented by the grid cell occupancy descriptors (GCODs) which are defined and discussed below in the Methods, and in the previous paper in this issue.[3] An in vivo VHTS should have an additional component to those of the in vitro VHTS. Descriptors reflecting transport/delivery and/or metabolism should be included in constructing the model screen. This is readily achieved in a 4D-QSAR analysis by adding appropriate descriptors, like log *P*,[4] to the trial descriptor pool used to build the 4D-QSAR model. If any of these non-GCOD descriptors are significant, they will survive in model optimization and appear in the resultant best 4D-QSAR models.

Exploration of QSAR model space is a critical component in constructing the optimum VHTS. A single, best QSAR model, in terms of some set of statistical measures of significance, may not be the best VHTS. This QSAR might not include compound features minimally sampled in the training set, but which are significant in governing the SAR of expanded compound libraries. Thus, it can be important to identify the *set* of unique QSAR models consistent with a training set to maximize the information available to build the VHTS. At issue is the inclusion/exclusion of descriptors which may be of minor statistical fitting significance for compounds of the training set, but which probe unique features of the compounds which become significant in the virtual screening of corresponding compound libraries.

† Current address: Pharmacopeia Inc., CN 5350, Princeton, NJ 08543.

10.1021/ci990032+ CCC: $18.00
Published on Web 09/18/1999

Table 1. Structure−Activity Data for the Glucose Analogue Inhibitors of Glycogen Phosphorylase Used in the 4D-QSAR Training Set



compd no.	α	β	K_i (mM)	ΔG_{303} (kcal/mol)
1	H	NHC(=O)CH_3	0.032	6.23
2	H	NHC(=O)CH_2CH_3	0.039	6.11
3	H	NHC(=O)CH_2Br	0.044	6.04
4	H	NHC(=O)CH_2Cl	0.045	6.03
5	H	NHC(=O)C_6H_5	0.081	5.67
6	H	NHC(=O)$CH_2CH_2CH_3$	0.094	5.58
7	H	NHC(=O)NH_2	0.14	5.34
8	H	C(=O)$NHCH_3$	0.16	5.26
9	H	NHC(=O)CH_2NH_2	0.37	4.76
10	C(=O)NH_2	H	0.37	4.76
11	H	C(=O)NH_2	0.44	4.65
12	H	C(=O)$NHNH_2$	0.40	4.17
13	H	SH	1.00	4.16
14	CH_2OH	H	1.50	3.92
15	OH	H	1.70	3.84
16	H	C(=O)NHC_6H_5	5.40	3.14
17	H	OH	7.40	2.95
18	H	CH_2CN	9.00	2.84
19	OH	CH_2OH	15.80	2.50
20	H	OCH_3	24.70	2.23
21	CH_2NH_2	H	34.50	2.03
22	C(=O)$NHCH_3$	H	36.70	1.99
23	CH_3	H	53.10	1.77
24	C(=O)NH_2	$NHCOOCH_3$	0.016	6.65
25	H	$NHCOOCH_2$Ph	0.35	4.79
26	H	NHC(=O)$CH_2NHCOCH_3$	0.99	4.17
27	H	C(=O)$NHNHCH_3$	1.8	3.81
28[a]	OH	H	2	3.74
29	H	C(=O)$NHCH_2CH_2OH$	2.6	3.58
30	H	$COOCH_3$	2.8	3.54
31	C(=O)$NHNH_2$	H	3.0	3.50
32	H	SCH_2C(=O)NHPh	3.6	3.39
33	H	C(=O)NH-4-OPh	4.4	3.27
34	H	$CH_2CH_2NH_2$	4.5	3.25
35	C(=O)NH-4-OHPh	H	5.6	3.12
36	OH	CH_2N_3	7.4	2.95
37	OH	CH_2CN	7.6	2.94
38	H	C(=O)$NHCH_2CH_2CF_3$	8.1	2.90
39	C(=O)NHPh	H	12.6	2.63
40	COOH	H	15.2	2.52
41	H	CH_2NH_2	16.8	2.46
42	C(=O)$NHCH_2CH_2OH$	H	16.9	2.46
43	H	SCH_2C(=O)NH-2,4-F_2Ph	18.9	2.39
44	H	SCH_2C(=O)NH_2	21.1	2.32
45	CH_2N_3	H	22.4	2.29
46	$COOCH_3$	H	24.2	2.24
47	C(=O)$NHCH_2$-2,4-F_2Ph	H	27.2	2.17

[a] The ring O replaced by S.

In this paper we describe a paradigm for applying 4D-QSAR analysis to construct effective VHTS models. The paradigm is applied to the evaluation of a focused virtual library of glycogen phosphorylase *b* (Gp*b*) inhibitors. The VHTS models are derived from the 4D-QSAR models constructed for a training set of Gp*b* inhibitors and reported in the preceding paper in this issue.[3]

METHODS

1. 4D-QSAR Models for Glucose Inhibitors of Gp*b*. The optimum 4D-QSAR model constructed from the training set in Table 1 for Gp*b* inhibition is given by eq 1 and shown in Figure 1. The alignment used to construct this 4D-QSAR model is shown in Figure 2.

$$\Delta G = 5.04\text{GC1(hbd)} - 2.68\text{GC2(np)} + 11.22\text{GC3(p}-\text{)} + 4.87\text{GC4(any)} + 2.76\text{GC5(p}+\text{)} - 1.35\text{GC6(any)} + 2.89 \quad (1)$$

$$N = 47 \qquad r^2 = 0.87 \qquad \text{xv-}r^2 = 0.83$$

In eq 1 ΔG is the binding free energy of an inhibitor to *Gpb*, GC1 through GC6 are the six grid cell occupancy

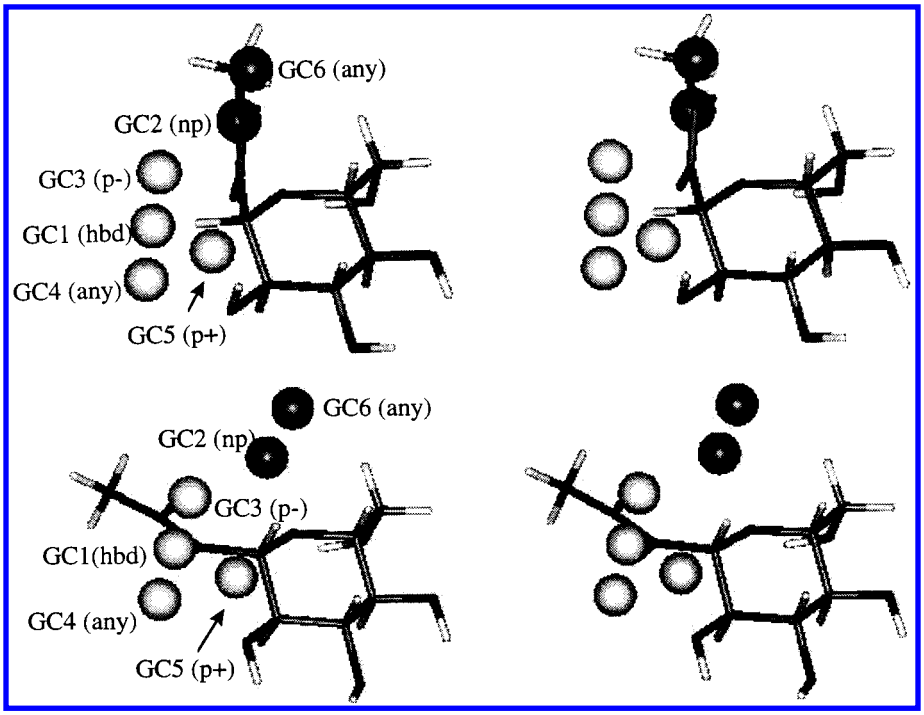
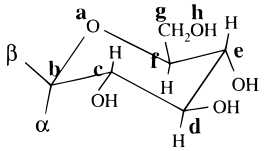


Figure 1. 4D-QSAR model given by eq 1 superimposed on (a, top) the weak inhibitor compound 22 of Table 1 and (b, bottom) the isomer of compound 22, compound 8 of Table 1, which is a moderately good binder. The spheres represent the GCODs given in eq 1. The light shaded GCODs define activity-enhancing GCODs, while the dark shaded GCODs are activity-decreasing sites. The molecules are represented as stick models in their respective postulated active conformations; see the preceding paper in this issue.



Alignment number	Atom 1	Atom 2	Atom 3
1	a	b	c

Figure 2. Alignment used to construct eq 1. The letters define the atoms used in defining the alignment sets.

descriptors, GCODs, found to be most significant for the training set of glucose inhibitor analogs given in Table 1, N is the number of inhibitors, r is the correlation coefficient, and xv-r is the cross-validation correlation coefficient. The symbol in parentheses next to each GCI refers to the type of atom occupying the grid cell which is significant to the QSAR. The symbol definitions are hbd = hydrogen bond donor, np = nonpolar, p− = polar negative charge density, any = any type of atom, and p+ = polar positive charge density. The GCODs define the locations in space, relative to the alignment, of key sites that should/should not be occupied by specific types of atoms and/or groups, generally referred to as *interaction pharmacophore elements* (IPEs), to optimize activity (ΔG). The spatial pharmacophore and steric site restrictions of the receptor for ligand binding are embedded in the 4D-QSAR model.

The details regarding the construction of eq l are in the preceding paper.[3] The grid cells are represented as spheres in Figure 1 where diameters are equal to the sides of the grid cells (1 Å for all grid cells). The dark colored spheres represent those grid cells whose occupancies by the requisite atom types, as given by eq 1, *decrease* inhibitor binding. Conversely, the light-colored spheres correspond to grid cells whose occupancy by a specific atom type *increases* ligand binding to Gp*b*. The identification code of each grid cell descriptor relative to eq 1 is also listed next to each sphere in Figure 1. The 4D-QSAR model is shown relative to (a) compound 22 ($\alpha = -C(=O)NHCH_3$, $\beta = -H$, $\Delta G = 1.99$ kcal/mol), a relatively poor inhibitor, and (b) its configurational isomer, compound 8 ($\alpha = -H$, $\beta = -C(=O)NHCH_3$, $G = 5.26$ kcal/mol), a good inhibitor. The compound number refers to the entries in Table 1, and the conformations shown in Figure 1 are the predicted "active" conformations from the 4D-QSAR analysis.

It is clear from Figure 1a that compound 22 is predicted to be a poor inhibitor because atoms of the $\alpha = -C(=O)NHCH_3$ group occupy the spaces of grid cells that decrease the binding free energy. Conversely, in Figure 1b it is seen that appropriate atoms of the $\beta = -C(=O)NHCH_3$ group occupy grid cells which enhance the binding free energy.

The non-GCOD descriptors log P, the octanol/water partition coefficient,[4] and molecular volume were included with the trial basis set of GCOD descriptors used to optimize the QSAR model. Obviously, neither of these descriptors survived in model optimization since they are not found in eq 1.

2. Construction of VHTS Models from 4D-QSAR Models. 4D-QSAR analysis normally leads to a highly oversolved problem in terms of statistical fitting. That is, the number of independent variables (GCODs) far exceeds the number of observations (compounds). This situation can be well-handled by a combination of data reduction (using partial least-squares, PLS, regressions[5]) and exhaustive model exploration and optimization (by genetic algorithm analysis[6] using the genetic function approximation, GFA[7]). The objective in QSAR model building is to maximize the statistical fit, based on some defined measure, of the dependent variables (ΔG) to the minimum number of

independent variables (GCODs). The lack of fit, LOF, measure,[8] employed to automate the GFA operation, is designed to achieve this objective of a maximum fit for a minimum number of independent variables. Thus, from the family of unique and significant 4D-QSAR models, that model with the fewest number of independent variables is selected as best.

However, a VHTS 4D-QSAR model, in addition to being reliable and accurate, should also explore the largest amount of *space* about the compounds in the training set. To be clear, this "space" is real three-dimensional space and not an abstract similarity/diversity space, derived from data reduction on a set of molecular descriptors, that is associated with the most current VHTS screens.[9−11] Maximum 3D spatial information from the training set is needed to make the corresponding VHTS relevant to the largest possible range in combinatorial structural diversity. The GCODs of 4D-QSAR analysis are well suited to optimize spatial information for a VHTS because these descriptors are derived from (a) sampling the thermodynamically relevant conformer states of each compound in the training set, (b) dividing each compound into the set of functional atom types associated with expression of biological activity, and (c) exploring the thermodynamic probability of atom-type occupancy at all sites in space containing each compound.

The building of the VHTS thus introduces the possible selection of GCODs, in addition to those in the statistically most significant 4D-QSAR models, as components of the virtual screen. Within the 4D-QSAR paradigm two different approaches to the selection of additional GCODs for VHTS models seem reasonable.

A. VHTS from the Manifold of 4D-QSAR Models. The set of significant, and independent, 4D-QSAR models, the *manifold* of 4D-QSAR models,[1,2] are used to construct a VHTS. The inclusion of a number of GCODs into a VHTS that is greater than permitted by overfitting constraints is "justified" because the GCODs are significant in the "smaller" independent 4D-QSAR models of the manifold. Two ways to build a VHTS from the manifold of 4D-QSAR models can be identified. One way to construct the VHTS is to use all the distinct GCODs of the manifold, and the ΔG values of the training set, to construct the corresponding regression equation. This equation is likely to be overfit from a statistical point of view, but may contain useful GCODs for virtual screening.

The other way to construct a VHTS from the manifold is to compute a *consensus* measure of ΔG from a *significance-weighted average* of all manifold models

$$\langle \Delta G[c] \rangle = \sum_i \Delta G_i[c] W_i \qquad (2)$$

where $\langle \Delta G[c] \rangle$ is the consensus ΔG of compound c, $\Delta G_i[c]$ is the estimated ΔG of c using the ith manifold 4D-QSAR model, and W_i is the relative *significance* of the ith model. Relative significance, W_i, can be defined in a number of arbitrary ways. One reasonable definition is the square of the correlation coefficient of model i weighted against the sum of the squares of the correlation coefficients of all the models in the manifold.

$$W_i = r_i^2 [\sum_j r_j^2]^{-1} \qquad (3)$$

The *stability* $S(c)$, of the consensus VHTS generated from the manifold of 4D-QSAR models is defined as the standard deviation of the consensus fits over the manifold

$$S(c) = [\sum_i (\langle \Delta G(c) \rangle - \Delta G_i(c) s_l)^2]^{1/2}/N \qquad (4)$$

In eq 4 N is the number of unique 4D-QSAR models in the manifold. A large value of $S(c)$, relative to the range in observed ΔG values, is indicative of a lack of stability in estimating $\Delta G(c)$ by consensus. That is, one or more of the manifold models is estimating a $\Delta G_i(c)$ value markedly different from the values of the other manifold models.

B. VHTS Using Group Enhancement. 4D-QSAR analyses can lead to a single 4D-QSAR model, as opposed to a manifold of models, which is the case for the training set of glucose inhibitors of Gp*b* in Table 1. The identification of a single 4D-QSAR model, eq 1 in this study, makes the interpretation of the findings from the 4D-QSAR analysis less ambiguous than when multiple models result from the analysis. However, the selection of the "best" GCODs, which overfit the 4D-QSAR model, but enhance the applicability of the resulting 4D-QSAR model as a VHTS, is ambiguous. In the most general case, increasingly large, in terms of the number of GCODs, 4D-QSAR models can be derived until overfitting produces no discernible increase in r^2 and/or xv-r^2. Any, or all, of the corresponding overfit 4D-QSAR models can be used as VHTS models.

However, our limited experience with 4D-QSAR models indicates that very often a model with $n + 1$ GCOD terms has n GCOD descriptors in common with an n GCOD term model. This is particularly true of 4D-QSAR models containing, or exceeding, the maximum number of statistically significant GCOD terms. This behavior is usually observed even when a manifold of unique 4D-QSAR models is found. Each unique model is the best model from a family of models containing common GCODs, but whose sizes, in terms of the number of GCODs, differ from one another.

These families of models can be represented in the following format:

$$\mathrm{DV} = a(1,1)\,\mathrm{GCOD}(1) + I(1) \qquad (5a)$$

$$r^2 = r^2(1)$$

$$\mathrm{DV} = a(1,2)\,\mathrm{GCOD}(1) + a(2,2)\,\mathrm{GCOD}(2) + I(2) \qquad (5b)$$

$$r^2 = r^2(1) + \Delta r^2(2)$$

$$\mathrm{DV} = a(1,3)\,\mathrm{GCOD}(1) + a(2,3)\,\mathrm{GCOD}(2) + a(3,3)\,\mathrm{GCOD}(3) + I(3) \qquad (5c)$$

$$r^2 = r^2(1) + \Delta r^2(2) + \Delta r^2(3)$$

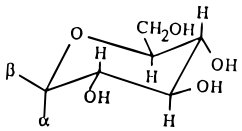
$$\vdots$$

$$\mathrm{DV} = {}^n\sum_{i=1} a(i,n)\,\mathrm{GCOD}(i) + I(n) \qquad (5d)$$

$$R^2 = r^2(1) + {}^n\sum_{n=2} \Delta r^2(i)$$

where DV is the dependent variable (ΔG), $a(i,n)$ is the regression coefficient of GCOD(i), $I(n)$ is the regression constant of fit, n is the number of GCODs in a model, r^2 is the square of the correlation coefficient, $r^2(1)$ is r^2 for a one-term GCOD model, and $\Delta r^2(n)$ is the nth enhancement of r^2 realized for a 4D-QSAR model having n GCOD terms as

Table 2. Substituents Used to Generate the Combinatorial Chemistry Virtual Library of 225 Glucose Analogue Inhibitors of Gp*b*



The set of 15 substituents for α and for β

1) H	6) $C(=O)CH_2$ - O - CH_3	11) CH_2 $C(=O)CH_2$ OH
2) $NHC(=O)CH_3$	7) CH_2 $C(=O)CH_3$	12) NO_2
3) $C(=O)NHCH_3$	8) $NHCH_3$	13) $NHC(=O)NH_2$
4) Cl	9) CH_2 $NHCH_3$	14) SO_2 NH_2
5) CH_3	10) CH_2 -O-CH_3	15) SH

compared to the $n - 1$ GCOD term model. Any other measure of statistical significance of fit besides r^2, such as xv-r^2, could also be used in this methodology.

The set of 4D-QSAR models represented by eqs 5a−d are called *additive enhancement (AE) models.* Any AE model with k GCOD terms can be viewed as "derived" from the corresponding AE model with $k - 1$ identical GCODs plus one new GCOD term and a data refitting by adjusting the GCOD regression coefficients. The number of GCOD terms (the upper value of n) retained in the largest model of an AE family of models can be determined by terminating n when $r^2(n) < \epsilon$ (ϵ being the limit on the increase in variance required to include the nth GCOD term). Again, other measures of fitting significance could also be employed, such as xv-r^2, to determine the terminal upper value, n, for the number of GCODs.

When a set of 4D-QSAR AE models are being explored as possible VHTSs, the contribution of spatial sampling of the AE model can also be considered in conjunction with measures of statistical significance of fit. One joint measure of statistical and spatial significance that might be used to define the size of n in eq 5d is

$$L(n) = \Delta r^2(n) P_n(\mathrm{GCOD}) \qquad (6)$$

P_n(GCOD) is the average atom-type occupancy of the nth term GCOD in eq 5d for all compounds in the training set. $L(n)$ can go to zero, which is the termination condition needed to define n in eq 5d, if either, or both, $\Delta r^2(n)$ and/or P_n(GCOD) go to zero. Thus, the statistical significance of the GCOD contributes to not only selection of n, but also the occupancy (spatial sampling) of this grid cell. Moreover, a low value of Δr^2 of a GCOD can be offset by a high average occupancy of the grid cell, and vice versa. A zero value of $L(n)$ indicates the nth candidate GCOD term does not contribute to the VHTS and the model is assumed to have converged with $n - 1$ terms.

Regardless of how a VHTS is constructed using the 4D-QSAR paradigm, it is applied to virtual library screening in the same way as any 4D-QSAR model for evaluating the activity of a test molecule. Each test molecule is assigned the alignment rule of the 4D-QSAR model, the GCOD values are correspondingly determined, and the activity is predicted from the model's regression equation [eq 1 in this case].

3. Selection of the Focused Virtual Glucose Analogue Library. The set of 15 substituents given in Table 2, representing moderate diversity from the training set of Table. 1, were selected for both the α and β subsitituent sites of the glucose core structure. Thus, a focused library of $15 \times 15 = 225$ virtual analogues was created. The analogues were built using the same procedure used for each of the compounds in the training set as described in ref 3. The GCODs of each of the analogs of the virtual library were constructed from a conformational ensemble profile, CEP, of 100 000 conformations using MDS and the same alignment used to construct eq 1. The small size of this library, 225 compounds, was chosen to conserve publication space. Still, it is important to provide some time estimates regarding the practical application limitations to VHTS using 4D-QSAR models. In this application a midlevel SGI O2 workstation was used to perform the VHTS. The processing time per library compound was 41.3 cpu s. Typical drug-candidate organic compounds (30−60 atoms) require in the range of 30−70 cpu s, per compound, to process. Processing time seems to scale in an inverse linear fashion with the processor speed of the workstation employed in the study.

The methodology given in this paper does not include a basis for selecting/constructing virtual libraries. However, one compound selection constraint is strongly encouraged. The library should be designed to include a few compounds from the training set. Moreover, the compounds selected from the training set should uniformly scan its activity range. The number of training set compounds built into a virtual library should be limited to a maximum of about 5% of the virtual library to maximize new SAR information, but also retain ties to the original training set. Five compounds from the training set of glucose inhibitors in Table 1 are embedded in the 225 compound virtual library.

Table 3. Normalized Significance Weights, W_i, of the GCOD Descriptors of Eq 1

no., *i*	GCOD	W_i	no., *i*	GCOD	W_i
1	GC1(hbd)	0.359	4	GC4(any)	0.042
2	GC2(np)	0.276	5	GC5(p+)	0.037
3	GC6(any)	0.263	6	GC3(p−)	0.022

RESULTS

The 225 analogs of the focused virtual library, defined by the substituents reported in Table 2, were screened using the 4D-QSAR model given by eq 1 as the VHTS. However, eq 1 was broken down into three, four, five, and six GCOD term representations to explore the AE model scheme for adding GCODs to define VHTS models. Table 3 contains the normalized significance weights, W_i, of each of the six GCODs of eq 1 computed using eq 3. From Table 3 it is seen that, for the training set, the first three GCODs of Table 3 have the highest significance weights and dominate the 4D-QSAR model. Thus, a three-term VHTS using these three GCODs can be considered the base-line model, which is the equivalent of eq 5a, for building and exploring AE models described in the Methods. It is important to remember that, like eq 1 itself, the W_i are dependent upon the training set. The set of W_i provide a reasonable basis for building larger (more GCODs) VHTS models in a consistent group additive fashion.

Table 4 reports the three, four, five, and six GCOD term AE VHTS models constructed from eq 1. The use of AE

Table 4. Three, Four, Five, and Six GCOD Term AE VHTS Models and Corresponding Statistical Measures Relative to the Training Set Given in Table 1[a]

(a) VHTS Models

	ranked GCOD no. from Table 3					
no. of GCODs	1	2	3	4	5	6
3	5:04	−2.68	−1.35			
4	5.04	−2.68	−1.35	4.87		
5	5.04	−2.68	−1.35	4.87	2.76	
6	5.04	−2.68	−1.35	4.87	2.76	−1.35

(b) Statistical Measures

no. of GCODs	r^2	Δr^2	no. of GCODs	r^2	Δr^2
3	0.79		5	0.84	0.01
4	0.83	0.04	6	0.87	0.03

[a] The regression constant of each VHTS is 2.89.

VHTS models derived from eq 1, which is the optimized 4D-QSAR model of the training set (Table 1), provides a controlled evaluation and analysis of a family of AE models. It is clear from Tables 3 and 4b that the first three GCOD descriptors account for the large majority of r^2. GCODs 4, 5, and 6 decreasingly provide an increasing contribution to r^2. Nevertheless, these three "minor" GCOD descriptors are judged by the GFA measure of fitting, the lack of fit, LOF, measure,[8] to provide discriminating information to predicting ΔG in the training set. Moreover, the identification and ranking of these six particular GCODs in constructing an optimum 4D-QSAR model represent a significant success in data reduction and model exploration. This successful data reduction and exploration can be illustrated using Figure 3. Figure 3a is a plot of grid cell number (*x*-axis) vs grid cell occupancy (*y*-axis) for the most active analogue of the training set, compound 24 of Table 1, for the 200 most highly weighted GCODs as found by PLS. Figure 3b is the same plot as Figure 3a, except the occupancy *difference* between compound 24 and the least active analogue, compound 23 of Table 1, is plotted on the *y*-axis.

Table 5 reports the predicted ΔG values for the 225 compounds of the focused virtual library using the three to six VHTS models given in Table 4. The compound coding $X-Y$ refers to substituent number X of Table 2 at the β substituent site and substituent number Y at the alpha substituent site. The 10 virtual compounds with the largest ΔG (tightest binding) values using the six-term VHTS are given in Table 6, and the 10 virtual compounds with the least ΔG (poorest binding) values are given in Table 7. Also listed in Table 7 are the five compounds common to the training set and the virtual library. Compound (2−14) [$\alpha = -SO_2NH_2$ and $\beta = -NHC(C{=}O)CH_3$] is predicted to be the best inhibitor in the virtual library and to bind with about 2 kcal/mol more ΔG than the best analogue, compound 24 of Table 1, in the training set (8.65 kcal/mol vs 6.65 kcal/mol). This additional binding free energy corresponds to a $K_i = 0.00059$ mM for (2−14) versus $K_i = 0.016$ mM for compound 24 of Table 1. The 10 best binding compounds of Table 6 are all predicted to bind better to Gp*b* than compound 24 of the training set.

The 10 poorest binders of the virtual library are all predicted to bind with a lower ΔG (see Table 7) than the poorest binder in the training set, compound 23 of Table 1 ($\alpha = -CH_3$, $\beta = -H$, $\Delta G = 1.77$ kcal/mol). The predicted ΔG of the poorest binding virtual compound (15−7) ($\alpha = -CH_2C(C{=}O)CH_3$, $\beta = -SH$) is 1.05 kcal/mol. Thus, the focused virtual library is predicted to spread the range from $\Delta G = 1.77-6.65$ kcal/mol for the training set to $\Delta G = 1.05-8.65$ kcal/mol. A histogram of the compound frequency distribution of the training set with respect to ΔG is shown in Figure 4a and for the focused virtual library in Figure 4b.

Virtual compounds predicted to have a ΔG at least as large (6.65 kcal/mol) as the best binder of the training set (compound 24 of Table 1) using the six-term VHTS were explored with respect to the GCOD source(s) of their ΔG values. Table 8 contains examples of good binding compounds from the virtual library which show marked increase, decrease, or no change in ΔG as a function of the number of GCODs used in the VHTS to estimate ΔG. The VHTSs are those reported in Table 4a. The $\Delta(\Delta G)$ in Table 8 refers to the change in the predicted ΔG in going from an *n*-term VHTS to an $(n-1)$-term VHTS, $\Delta[n-(n-1)]$ defines the specific pair of AE VHTS models. In this particular application, the inclusion of four and/or five GCODs in the AE VHTS models always increases the predicted value of ΔG. This correlates with the positive regression coefficients of GCODs 4 and 5 of Table 4a. Conversely, the addition of the sixth GCOD of Table 4a corresponds to either no change or a modest decrease in the predicted ΔG relative to the five-term VHTS. The regression coefficient of the sixth GCOD of Table 4a is negative so that the loss in predicted ΔG values again correlates directly with the sign of the regression coefficient.

The direct correlation of increases/decreases in ΔG with the corresponding sign of the GCOD regression coefficient suggests these are site-additive 4D-QSAR models with little coupling, or allosteric character, among the GCODs. That is, the gain in GCOD occupancy at one site to increase/decrease ΔG is not offset by a corresponding change in GCOD occupancy at another site to decrease/increase ΔG. This uncoupled behavior is different from the coupled (allosteric) behavior found among GCODs in the 4D-QSAR models for benzylpyrimidine inhibitors of dihydrofolate reductase[1] and interphenylene 7-oxabicycloheptane oxazole thromboxane A_2 receptor antagonists.[2]

DISCUSSION

A dilemma that seems to have gone largely untreated in performing VHTS is estimating how appropriate a model from a particular training set is for the evaluation of a particular virtual library. An *implicit assumption* seems to be that if a large enough basis set of (predominantly 2D) descriptors is derived from the training set, the resultant VHTS from data-reduction of that descriptor pool will be "appropriate" to apply to an arbitrary virtual library. Moreover, when 3D descriptors are included in the trial basis set of descriptors, the conformations and alignments used in descriptor estimation are also *assumed* to be appropriate. Little concern seems to be given to the fact that a poor selection of conformation and/or alignment can actually lead to a loss in the quality of VHTS relative to excluding these descriptors from the trial basis set.

The properties of a GCOD make possible the explicit estimation of how well a given 4D-QSAR model employed

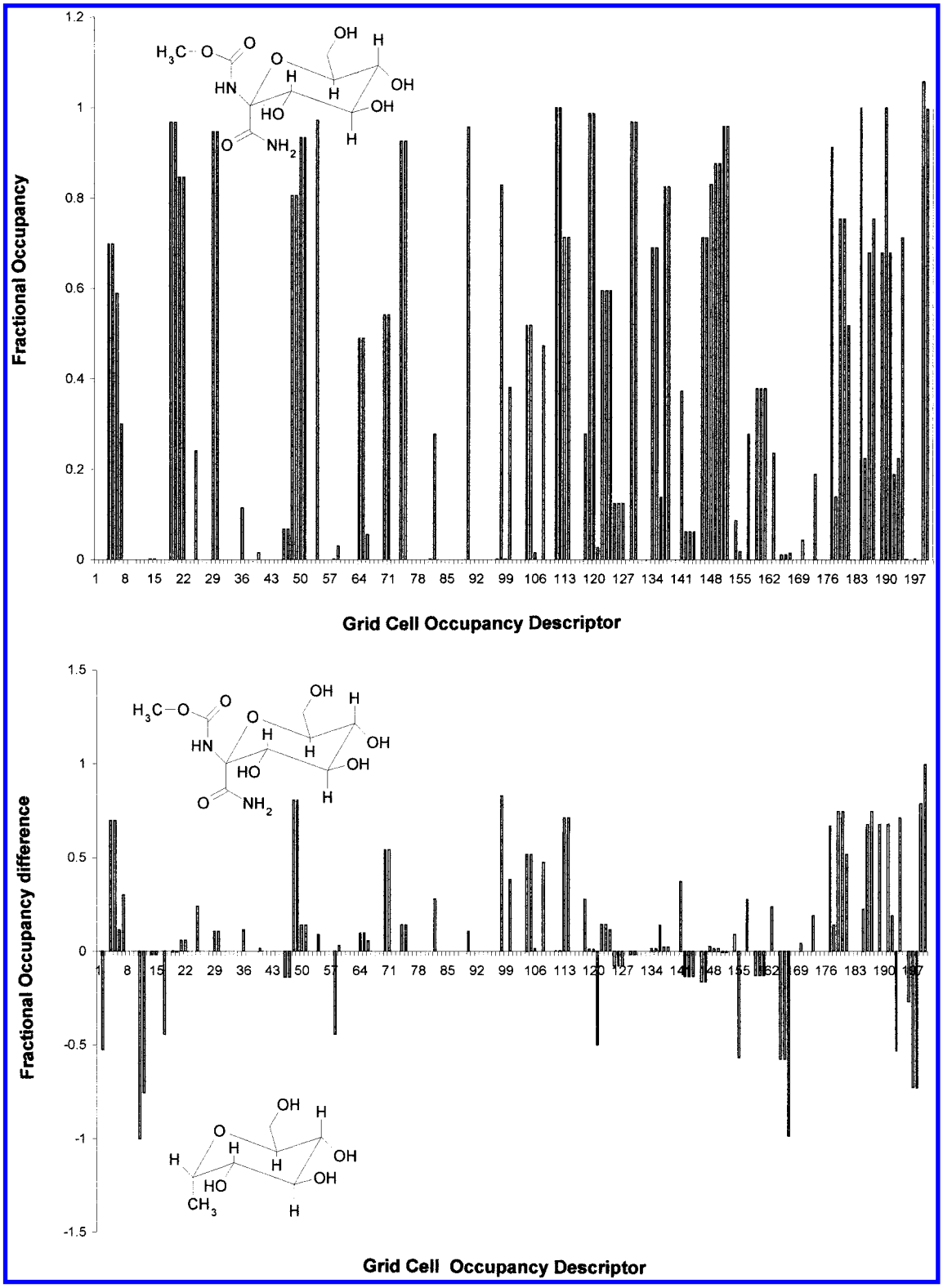


Figure 3. The 200 most highly (no. 1 the highest) weighted PLS GCODs vs (a, top) the occupancy of the GCODs for the best binding compound (24 of Table 1) and (b, bottom) the occupancy difference of compound 24 and compound 23, the weakest binding compound of the training set.

as a VHTS explores, and evaluates, a given virtual library. Each GCOD defines a unique, and orthogonal, position in space relative to the other GCODs and alignment criteria. Moreover, each atom type (IPE) is different (unique) from all other atom types (including the *all* atom type IPE). Thus, each GCOD defines a "piece of information" unique (orthogonal) from all other GCODs.

A comparison of usage/occupancy of the GCODs of a virtual library processed through a VHTS to the GCODs of the training set used to derive the VHTS permits a direct measure of how appropriate the VHTS is for the virtual library. A 4D-QSAR model can predict activities beyond the range of the training set without having to explore "binding space" beyond that sampled by the training set. The

Table 5. Predicted ΔG of Each Compound of the Focused Virtual Library Using the Three to Six-Term VHTS Models Given in Table 4[a]

	predicted ΔG values					predicted ΔG values					predicted ΔG values			
compd $X-Y$	six-term	five-term	four-term	three-term	compd $X-Y$	six-term	five-term	four-term	three-term	compd $X-Y$	six-term	five-term	four-term	three-term
1-1	2.89	2.89	2.89	2.89	14-10	2.06	2.89	2.89	2.89	5-10	1.91	1.91	1.91	1.91
1-10	2.04	2.04	2.04	2.04	14-11	2.04	2.04	2.04	2.04	5-11	7.86	7.86	7.86	2.66
1-11	2.00	2.00	2.00	2.00	14-12	2.89	2.89	2.89	2.89	5-12	2.89	2.89	2.89	2.89
1-12	2.89	2.89	2.89	2.89	14-13	2.87	2.88	2.88	2.88	5-13	2.79	2.80	2.80	2.80
1-13	2.89	2.89	2.89	2.89	14-14	2.60	2.89	2.89	2.89	5-14	2.89	2.89	2.89	2.89
1-14	2.99	3.04	2.93	2.89	14-15	2.36	2.36	2.36	2.36	5-15	2.89	2.89	2.89	2.89
1-15	2.89	2.89	2.89	2.89	14-2	2.85	2.86	2.86	2.86	5-2	2.88	2.89	2.89	2.89
1-2	2.89	2.90	2.89	2.89	14-3	2.42	2.89	2.89	2.89	5-3	2.91	2.91	2.91	2.89
1-3	3.22	3.23	3.23	2.88	14-4	2.90	2.90	2.90	2.90	5-4	2.89	2.89	2.89	2.89
1-4	2.89	2.89	2.89	2.89	14-5	1.91	1.91	1.91	1.91	5-5	1.81	1.81	1.81	1.81
1-5	2.01	2.01	2.01	2.01	14-6	2.14	2.63	2.63	2.63	5-6	2.44	2.51	2.51	2.47
1-6	4.03	4.06	4.06	2.62	14-7	1.75	2.08	2.08	2.08	5-7	2.06	2.06	2.06	2.06
1-7	2.11	2.12	2.12	2.12	14-8	2.46	2.47	2.47	2.47	5-8	2.89	2.89	2.89	2.89
1-8	2.89	2.89	2.89	2.89	14-9	1.81	2.89	2.89	2.89	5-9	2.22	2.22	2.22	2.22
1-9	1.92	1.92	1.92	1.92	15-1	2.89	2.89	2.89	2.89	6-1	5.53	5.53	5.53	5.17
10-1	2.89	2.89	2.89	2.89	15-10	1.69	1.69	1.69	1.69	6-10	5.21	5.21	5.21	5.21
10-10	1.78	1.78	1.78	1.78	15-11	1.63	1.64	1.64	1.64	6-11	4.93	4.93	4.93	4.93
10-11	1.17	1.17	1.16	1.16	15-12	2.89	2.89	2.89	2.89	6-12	6.06	6.06	6.06	6.05
10-12	2.89	2.89	2.89	2.89	15-13	2.95	2.95	2.91	2.89	6-13	6.29	6.29	6.29	6.28
10-13	2.88	2.91	2.91	2.89	15-14	3.05	3.05	3.05	3.05	6-14	6.27	6.27	4.87	4.85
10-14	2.89	2.89	2.89	2.89	15-15	1.96	1.96	1.96	1.96	6-15	5.60	5.60	5.60	5.60
10-15	1.24	1.24	1.24	1.24	15-2	3.08	3.08	3.08	3.07	6-2	6.31	6.31	6.31	6.30
10-2	2.89	2.91	2.91	2.89	15-3	2.80	3.04	3.01	2.99	6-3	5.67	6.04	5.95	5.92
10-3	2.84	3.10	3.07	2.89	15-4	2.89	2.89	2.89	2.89	6-4	5.86	5.86	5.86	5.86
10-4	2.89	2.89	2.89	2.89	15-5	1.86	1.86	1.86	1.86	6-5	5.05	5.05	5.05	5.04
10-5	1.86	1.86	1.86	1.86	15-6	2.61	2.71	2.70	2.56	6-6	6.16	6.16	6.16	5.88
10-6	2.69	2.73	2.73	2.66	15-7	1.05	1.05	1.05	0.98	6-7	5.15	5.15	5.15	5.15
10-7	1.18	1.18	1.18	1.18	15-8	2.98	2.98	2.98	2.89	6-8	6.47	6.47	6.47	6.19
10-8	2.98	2.98	2.98	2.89	15-9	1.34	1.34	1.34	1.31	6-9	4.65	4.65	4.65	4.47
10-9	1.34	1.34	1.34	1.28	2-1	7.17	7.17	7.17	7.10	7-1	2.88	2.88	2.88	2.86
11-1	2.90	2.90	2.90	2.85	2-10	6.35	6.35	6.35	6.35	7-10	1.82	1.82	1.82	1.82
11-10	1.87	1.87	1.87	1.79	2-11	6.07	6.07	6.07	6.07	7-11	1.27	1.27	1.27	1.27
11-12	3.18	3.18	3.18	2.89	2-12	7.27	7.27	7.27	7.26	7-12	2.89	2.89	2.89	2.89
11-13	2.90	2.95.	2.95	2.89	2-13	7.16	7.16	7.16	7.16	7-13	2.89	2.89	2.89	2.89
11-14	2.51	2.92	2.92	2.89	2-14	8.65	8.65	7.41	7.41	7-14	2.47	2.89	2.89	2.89
11-15	2.65	2.65	2.65	2.62	2-15	3.45	3.45	3.45	3.44	7-15	2.76	2.76	2.76	2.76
11-2	2.92	2.92	2.92	2.89	2-2	7.35	7.35	7.35	7.35	7-2	2.88	2.89	2.89	2.89
11-3	4.68	4.69	2.98	2.89	2-3	8.08	8.24	7.43	7.29	7-3	4.30	4.34	3.09	2.89
11-4	2.90	2.90	2.90	2.89	2-4	7.17	7.17	7.17	7.17	7-4	2.89	2.89	2.89	2.89
11-5	1.91	1.91	1.91	1.89	2-5	6.31	6.31	6.31	6.30	7-5	2.05	2.05	2.05	2.05
11-6	3.03	3.03	3.03	2.89	2-6	7.25	7.27	7.27	7.27	7-6	2.88	2.89	2.89	2.86
11-7	1.26	1.26	1.26	1.25	2-7	6.05	6.05	6.05	6.05	7-7	1.21	1.21	1.21	1.18
11-8	2.90	2.90	2.90	2.89	2-8	7.33	7.33	7.33	7.31	7-8	2.89	2.89	2.89	2.89
11-9	1.34	1.34	1.34	1.30	2-9	8.09	8.09	5.86	5.86	7-9	1.36	1.36	1.36	1.35
12-1	2.96	2.96	2.96	2.96	3-1	4.53	4.53	4.53	4.36	8-1	2.95	2.95	2.95	2.89
12-10	2.03	2.03	2.03	2.03	3-10	5.04	5.04	5.04	5.03	8-10	1.84	1.84	1.84	1.83
12-11	2.28	2.28	2.28	2.28	3-11	4.36	4.36	4.36	4.36	8-11	1.27	1.27	1.27	1.27
12-12	3.21	3.21	3.21	3.21	3-12	5.36	5.36	5.36	5.25	8-12	2.90	2.90	2.90	2.89
12-13	4.02	4.02	4.02	4.01	3-13	5.03	5.03	5.03	5.00	8-13	2.87	2.90	2.90	2.89
12-14	2.82	3.10	3.10	3.10	3-14	5.21	5.21	5.19	5.17	8-14	2.89	2.89	2.89	2.89
12-15	2.24	2.24	2.24	2.24	3-15	4.69	4.69	4.69	4.69	8-15	2.77	2.77	2.77	2.76
12-2	3.32	3.32	3.32	3.32	3-2	5.51	5.51	5.51	5.41	8-2	2.90	2.90	2.90	2.89
12-3	4.73	4.92	4.64	3.03	3-3	7.41	7.45	5.82	5.80	8-3	4.33	4.34	2.95	2.89
12-4	3.15	3.15	3.15	3.15	3-4	5.68	5.68	5.68	5.68	8-4	2.89	2.89	2.89	2.89
12-5	2.16	2.16	2.16	2.16	3-5	4.24	4.24	4.24	4.24	8-5	1.88	1.88	1.88	1.83
12-6	2.63	2.68	2.68	2.67	3-6	5.23	5.25	5.25	5.24	8-6	3.03	3.03	3.03	2.89
12-7	2.39	2.39	2.39	2.39	3-7	4.36	4.36	4.36	4.36	8-7	1.23	1.23	1.23	1.23
12-8	3.76	3.76	3.76	3.68	3-8	5.29	5.29	5.29	5.25	8-8	2.95	2.95	2.95	2.89
12-9	1.70	1.70	1.70	1.66	3-9	4.58	4.58	4.58	4.58	8-9	1.52	1.52	1.52	1.46
13-1	7.08	7.08	7.08	7.03	4-1	2.89	2.89	2.89	2.89	9-1	2.89	2.89	2.89	2.89
13-10	6.29	6.29	6.29	6.28	4-10	2.02	2.02	2.02	2.02	9-10	1.90	1.90	1.90	1.90
13-11	6.56	6.56	6.56	6.56	4-11	2.05	2.05	2.05	2.05	9-11	1.19	1.19	1.19	1.19
13-12	7.36	7.36	7.36	7.18	4-12	2.89	2.89	2.89	2.89	9-12	2.89	2.89	2.89	2.89
13-13	5.25	5.25	5.25	5.23	4-13	2.77	2.79	2.79	2.79	9-13	2.89	2.89	2.89	2.89
13-14	7.30	7.30	7.30	7.25	4-14	2.80	2.89	2.89	2.89	9-14	2.51	2.89	2.89	2.89
13-15	7.44	7.44	7.44	7.42	4-15	2.89	2.89	2.89	2.89	9-15	1.55	1.55	1.55	1.55
13-2	6.28	6.28	6.28	6.27	4-2	2.40	2.40	2.40	2.40	9-2	2.89	2.89	2.89	2.89
13-3	6.99	7.32	7.31	7.26	4-3	2.82	2.89	2.89	2.89	9-3	3.49	3.58	3.25	2.89
13-4	7.48	7.48	7.48	7.43	4-4	2.89	2.89	2.89	2.89	9-4	2.89	2.89	2.89	2.89
13-5	6.38	6.38	6.38	6.37	4-5	1.92	1.92	1.92	1.92	9-5	1.87	1.87	1.87	1.87
13-6	7.23	7.25	7.25	7.20	4-6	2.19	2.25	2.25	2.25	9-6	2.93	2.94	2.94	2.86
13-7	6.36	6.36	6.36	6.34	4-7	2.00	2.00	2.00	2.00	9-7	1.21	1.21	1.21	1.21
13-8	7.18	7.18	7.18	7.15	4-8	2.89	2.89	2.89	2.89	9-8	2.97	2.97	2.97	2.89
13-9	6.32	6.32	6.32	6.32	4-9	1.89	1.89	1.89	1.89	9-9	1.35	1.35	1.35	1.31
14-1	2.94	2.94	2.94	2.89	5-1	2.89	2.89	2.89	2.89	11-11	1.25	1.25	1.25	1.24

[a] Compound $X-Y$ refers to substituent number X from Table 2 on the β site and number Y on the α site.

Table 6. The 10 Most Active [Virtual] Analogues of the [Virtual] Combinatorial Chemistry Library of 225 Analogues from the [Virtual] HTS

rank	α	β	pred K_i (mM)	ΔG (kcal/mol)
1	SO_2NH_2	NHC(=O)CH_3	0.00059	8.65
2	CH_2NHCH_3	NHC(=O)CH_3	0.0015	8.09
3	C(=O)$NHCH_3$	NHC(=O)CH_3	0.0017	8.08
4	CH_2C=OCH_2OH	CH_3	0.0022	7.86
5	CH_3	NHC(=O)NH_2	0.0041	7.48
6	SH	NHC(=O)NH_2	0.0050	7.43
7	C(=O)$NHCH_3$	C(=O)$NHCH_3$	0.0056	7.40
8	NO_2	NHC(=O)NH_2	0.0068	7.35
9	NHC(=O)CH_3	NHC(=O)CH_3	0.0072	7.34
10	$NHCH_3$	NHC(=O)CH_3	0.0076	7.33
most active ligand in the training set	C(=O)NH_2	NHC(=O)OCH_3	0.016	6.65

Table 7.

(a) The 10 Least Active [Virtual] Analogues of the [Virtual] Combinatorial Chemistry Library of 225 Analogues from the [Virtual] HTS

rank	α	β	pred K_i (mM)	ΔG (kcal/mol)
1	CH_2C(=O)CH_3	SH	173	1.05
2	CH_2C=OCH_2OH	CH_2OCH_3	142	1.17
3	CH_2(C=O)CH3	CH_2OCH_3	138	1.19
4	SH	CH_2OCH_3	126	1.24
5	CH_2C(=O)CH_3	CH_2C=OCH_2OH	122	1.26
6	CH_2NHCH_3	CH_2OCH_3	114	1.34
7	CH_2NHCH_3	CH_2C=OCH_2OH	114	1.34
8	CH_2NHCH_3	SH	114	1.34
9	CH_2C=OCH_2OH	CH_2C(=O)CH_3	108	1.35
10	C(=O)C_2OCH_3	CH_2(C=O)CH_3	103	1.36
least active ligand in the training set	CH_3	H	53.1	1.77

(b) The 5 Analogues of the Virtual Combinatorial Chemistry Library of 225 Analogues Which Are Common to the VHTS Training Set

no.	α	β	pred ΔG (kcal/mol)	obsd ΔG (kcal/mol)
1	H	NHC(=O)CH_3	7.17	6.23
2	H	C(=O)$NHCH_3$	4.82	5.26
3	C(=O)$NHCH_3$	H	2.22	1.99
4	H	SH	3.89	4.16
5	H	CH_3	2.00	1.77

occupancy property of the GCODs of a 4D-QSAR model permits higher and lower activities (relative to the training set) to be predicted for virtual library compounds that do not reach beyond the space sampled by the training set.

Moreover, the capacity to develop overfit VHTS models, in which the descriptors (GCODs) are spatially orthogonal to one another, permits a VHTS to be extended to better explore a particular virtual library. Highly sampled GCODs of the virtual library that are poorly sampled in the training set can be readily identified and corresponding compounds flagged as sampling new possible ligand−receptor spaces.

The VHTS methodology using 4D-QSAR analysis currently does not include techniques to design a virtual library, nor does the VHTS methodology using 4D-QSAR analysis allow a structure optimization as a function of the end point of the screen. Our work to date has focused on developing the best VHTS from the information provided in a 4D-QSAR analysis. However, an analysis of the predictions from a VHTS can provide *patterns* of structure−activity information useful in new compound design. For example, in the case of the VHTS of the 225 virtual glucose analogues the following structure−activity patterns emerge from inspections of Tables 6 and 7a: (1) An −NH− group beginning (from the ring) a β-substituent is good for binding. (2) A $-CH_2-$ group beginning (from the ring) a β-substituent is bad for binding. (3) More chemical structure variability in a substituent is allowed at the α site than at the β site.

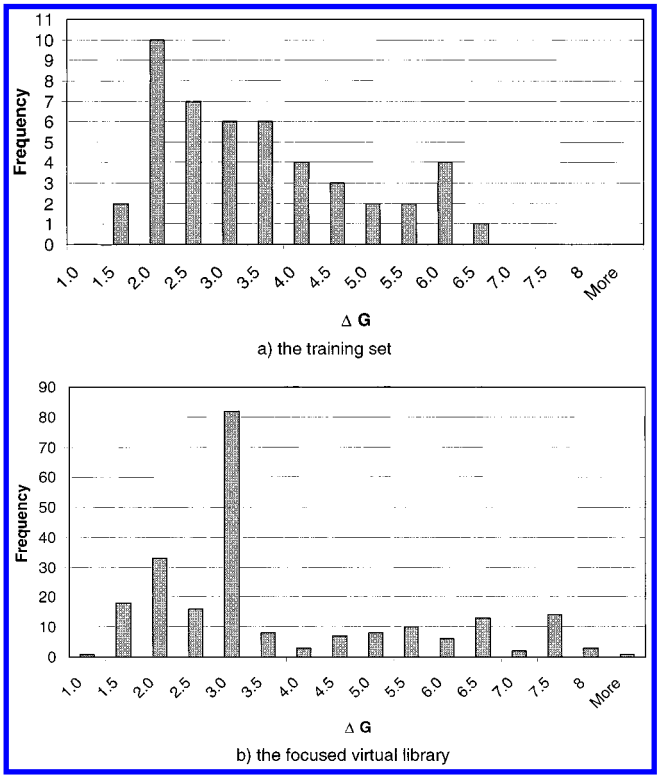


Figure 4. Histograms of ΔG (x-axis) vs compound frequency for (a) the training set and (b) the virtual library.

Table 8. Examples of Virtual Compounds Predicted To Be Good Binders Using the Six-Term VHTS of Table 4a Which Demonstrate Increases, Decreases, or No Change in ΔG as a Function of the Number of GCODs in the Additive Enhancement VHTS As Defined in Table 4a

		$\Delta(\Delta G)$		
compd *X*−*Y*	ΔG six-term	Δ(6-5)	Δ(5-4)	Δ(4-3)
5-11	7.86	0.00	0.00	**5.20**
2-9	8.09	0.00	**2.23**	0.00
2-6	7.25	**−0.02**	0.00	0.00
2-2	7.35	0.00	0.00	0.00
2-3	8.08	**−0.16**	**0.81**	**0.14**

A medicinal chemist might infer one or more of these structure−activity patterns from the training set of compounds and corresponding activities. However, the confidence of an investigator that these patterns are significant will be limited because of the very limited number of examples in the training set data. Moreover, other possible patterns of structure−activity can also be inferred from the training set of data. These spurious patterns are eliminated, and the significant patterns of structure−activity are amplified by a 4D-QSAR VHTS. Still, it is the ability to explicitly design focused libraries which is the primary application use of the 4D-QSAR VHTS approach and remains the focus of our current work.

ACKNOWLEDGMENT

This work was supported, in part, from resources of the Laboratory of Molecular Modeling and Design at the University of Illinois at Chicago. We also acknowledge the financial support of The Chem21 Group, Inc.

REFERENCES AND NOTES

(1) 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.

(2) 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 A_2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−938.

(3) 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.

(4) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

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

(6) 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.

(7) 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.

(8) Friedman, J. Multivariate adaptive regression splines. Technical Report No. 102; Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford, CA, Nov 1988 (revised Aug 1990).

(9) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New perspectives in lead generation II: evaluating molecular diversity. *Drug Discovery Today* **1996**, *1*, 71−78.

(10) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: using MDL keys as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(11) Turner, D. B.; Tyrell, S. M.; Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.

CI990032+