Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase

Dahua Pan, Yufeng Tseng, and A. J. Hopfinger*

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 15, 2003

A method for performing quantitative structure-based design has been developed by extending the current receptor-independent RI-4D-QSAR methodology to include receptor geometry. The resultant receptor-dependent RD-4D-QSAR approach employs a novel receptor-pruning technique to permit effective processing of ligands with the lining of the binding site wrapped about them. Data reduction, QSAR model construction, and identification of possible pharmacophore sites are achieved by a three-step statistical analysis consisting of genetic algorithm optimization followed by backward elimination multidimensional regression and ending with another genetic algorithm optimization. The RD-4D-QSAR method is applied to a series of glucose inhibitors of glycogen phosphorylase b, GPb. The statistical quality of the best RI- and RD-4D-QSAR models are about the same. However, the predictivity of the RD- model is quite superior to that of the RI-4D-QSAR model for a test set. The superior predictive performance of the RD- model is due to its dependence on receptor geometry. There is a unique induced-fit between each inhibitor and the GPb binding site. This induced-fit results in the side chain of Asn-284 serving as both a hydrogen bond acceptor and donor site depending upon inhibitor structure. The RD-4D-QSAR model strongly suggests that quantitative structure-based design cannot be successful unless the receptor is allowed to be completely flexible.

INTRODUCTION

Structure-based design, SBD, currently involves using a 3D-structure for the receptor, or better, a 3D-structure of a ligand−receptor complex, to perform molecular modeling studies to elucidate the features of ligand−receptor binding and virtually screen untested ligands. The 3D-structure of the receptor comes from X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and/or homology model building. Most current SBD approaches restrict the conformational freedom of the receptor geometry to some extent, often assuming a completely rigid geometry. Because of these geometric constraints, shortcomings in force fields and the neglect of some of the ligand−receptor binding contributions, even the semiquantitative prediction of ligand−receptor thermodynamics has remained problematic. That is, the computational equivalent of the in vitro binding assay has remained an elusive goal. SBD prediction of high affinity ligands is particularly difficult and unreliable.

We have developed a combination QSAR and SBD methodology[1−3] called free energy force field (FEFF) 3D-QSAR analysis which permits performing reliable quantitative SBD. Computational in vitro binding analyses (virtual screens) can be performed using FEFF 3D-QSAR analysis. FEFF 3D-QSAR analysis achieves quantitative SBD capability by (1) including all thermodynamic contributions to ligand−receptor binding in a (aqueous) solvent medium, (2) using a training set, normally consisting of a common receptor and a set of known (analog) ligands, to train the ligand−receptor force field, and (3) using the trained ligand−receptor force field, which is a QSAR, to virtually screen ligands outside of the training set. Thus, the FEFF 3D-QSAR works by calibrating a force field to fit the binding thermodynamics of a given ligand−receptor system. The calibrated force field is only valid for the ligand−receptor system from which it is derived.

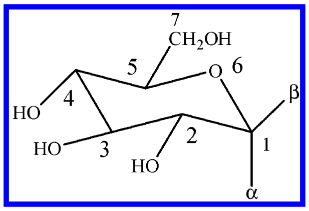
A drawback to the FEFF 3D-QSAR models is that they are composed completely of thermodynamic/force field terms as the descriptor. These descriptors do not permit any visual or geometric insight regarding ligand−receptor binding. No 3D-pharmacophore information is provided in a FEFF 3D-QSAR model. Thus, hypothesis generation regarding both lead optimization and lead identification are each difficult to ascertain from FEFF 3D-QSAR models.

We have also developed 4D-QSAR analysis[4−10] and originally applied it to structure−activity data sets where, in each case, the geometry of the receptor is *NOT* available. This form of 4D-QSAR analysis is defined as receptor-independent RI-4D-QSAR analysis. 4D-QSARs are quantitative 3D-pharmacophore, precisely the type of models we would like to have to complement FEFF 3D-QSAR models. Moreover, a methodology to determine how and to what extent the receptor contributes to each ligand that binds to it would be useful in ligand design. The ability to meaningfully compare a receptor-independent (RI) ligand−receptor binding model and to extract explicit receptor binding information, requires both models (RI and RD) to be constructed using a common paradigm and corresponding formalism. Thus, it would be advantageous to have a RD-4D-QSAR

* Corresponding author phone: (312)996-4816; e-mail: hopfingr@uic.edu.

10.1021/ci0340714 CCC: $25.00
Published on Web 08/05/2003

Table 1. Structure−Activity Data of the Training Set GPb Inhibitors Used in the 4D-QSAR Analyses



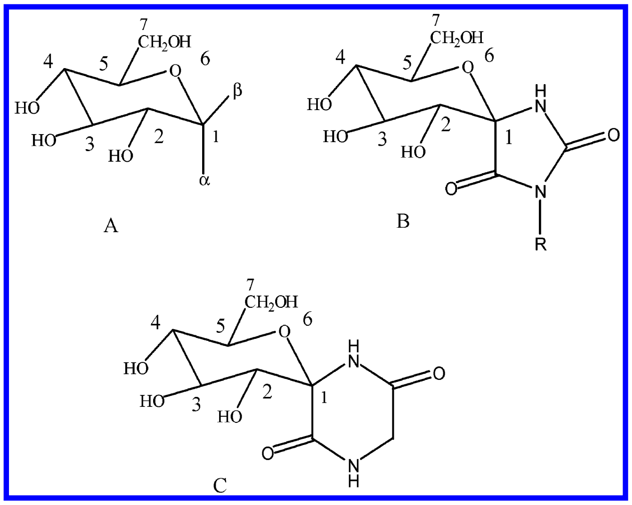
compd no.	α	*β*	K_i (mM)	Δ*G*303 (kcal/mol)
1	H	$NHCOCH_3$	0.032	6.23
2	H	$NHCOCH_2CH_3$	0.039	6.11
3	H	$NHCOCH_2Br$	0.044	6.04
4	H	$NHCOCH_2Cl$	0.045	6.03
5	H	$NHCOC_6H_5$	0.081	5.67
6	H	$NHCOCH_2CH_2CH_3$	0.094	5.58
7	H	$NHCONH_2$	0.14	5.34
8	H	$CONHCH_3$	0.16	5.26
9	H	$NHCOCH_2NH_3^+$	0.37	4.76
10	$CONH_2$	H	0.37	4.76
11	H	$CONH_2$	0.44	4.65
12	H	$CONHNH_2$	0.4	4.17
13	H	SH	1	4.16
14	CH_2OH	H	1.5	3.92
15	OH	H	1.7	3.84
16	H	$CONHC_6H_5$	5.4	3.14
17	H	OH	7.4	2.95
18	H	CH_2CN	9	2.84
19	OH	CH_2OH	15.8	2.5
20	H	OCH_3	24.7	2.23
21	$CH_2NH_3^+$	H	34.5	2.03
22	$CONHCH_3$	H	36.7	1.99
23	CH_3	H	53.1	1.77
24	$CONH_2$	$NHCOOCH_3$	0.016	6.65
25	H	$NHCOOCH_2Ph$	0.35	4.79
26	H	$NHCOCH_2NHCOCH_3$	0.99	4.17
27	H	$CONHNHCH_3$	1.8	3.81
28[a]	OH	H	2	3.74
29	H	$CONHCH_2CH_2OH$	2.6	3.58
30	H	$COOCH_3$	2.8	3.54
31	$CONHNH_2$	H	3	3.5
32	H	$SCH_2CONHPh$	3.6	3.39
33	H	CONH-4-OHPh	4.4	3.27
34	H	$CH_2CH_2NH_3^+$	4.5	3.25
35	CONH-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	$CONHCH_2CF_3$	8.1	2.9
39	CONHPh	H	12.6	2.63
40	COO^-	H	15.2	2.52
41	H	$CH_2NH_3^+$	16.8	2.46
42	$CONHCH_2CH_2OH$	H	16.9	2.46
43	H	SCH_2CONH-2,4-F_2Ph	18.9	2.39
44	H	SCH_2CONH_2	21.1	2.32
45	CH_2N_3	H	22.4	2.29
46	$COOCH_3$	H	24.2	2.24
47	$CONHCH_2$-2,4-F_2Ph	H	27.2	2.17

[a] The O on glucose ring is replaced by S.

methodology. Moreover, RD-4D-QSAR analysis would be the bridge connecting RI-4D-QSAR and FEFF-3D-QSAR. The three QSAR methodologies, in composite, would constitute a self-consistent manifold approach for building comparable QSAR models based on varying available level of structure−activity information.

This paper reports the RD-4D-QSAR methodology and the application of the methodology to a series of glucose analogue inhibitors of glycogen phosphorylase.

Table 2. Structure−Activity Data of the Test Set GPb Inhibitors Used in the 4D-QSAR Analyses



compd no.	structure	substituents	K_i (mM)	Δ*G* (kcal/mol)
T1	A	α-H, *β*-$CH_2OSO_2CH_3$	4.8	3.21
T2	A	α-OH, *β*-$CH_2OSO_2CH_3$	3.7	3.37
T3	A	α-H, *β*-1H-indol-2yloxy	2.6	3.58
T4	A	α-H, *β*-CONH-cyclopropyl	1.3	4
T5	A	α-H, *β*-$NHCOOCH_3$	0.085	5.64
T6	B	NH_2	0.146	5.32
T7	B	H	0.003	7.66
T8	C		0.0597	5.86

MATERIALS

Training Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase-b (GPb). The structure−activity data of the 47 glucose analogue inhibitors forming the training set used to construct both the receptor-independent (RI) and receptor-dependent (RD) models are given in Table 1. These structures, and the corresponding inhibitory binding constants (K_i), are reported in refs 11−14. The free energy of binding, Δ*G*, is estimated from K_i by

$$\Delta G = -RT \ln K_i \qquad (1)$$

where *T* is the absolute temperature and *R* is the gas constant. Δ*G* is used as the measure of the biological response for constructing the RD- and RI-4D-QSAR models.

Test Set of Glucose Analogue Inhibitors of GPb. The structure−activity data of the eight analogues in Table 2 were used as a test set. These structures, together their inhibition binding constants, are taken from ref 14. The Δ*G* of binding for the test set inhibitors were also calculated using eq 1.

The Crystal Structure of Glucose-Bound Receptor Complex. The X-ray crystal structure of the T-state of glucose-bound glycogen phosphorylase-b was determined by Martin and co-workers[15] to a resolution of 2.3 Å. The corresponding PDB file (2GPB) was obtained from the Brookhaven Protein Data Bank.[16] An illustration of the structure of the protein-glucose inhibitor is shown in Figure 1 where it seen that the bound glucose inhibitor is buried 12 Å within the protein. Water molecules identified in the crystal structure have been excluded for the 4D-QSAR analyses.

RD-4D-QSAR ANALYSIS

The details of the RI-4D-QSAR formalism has been reported by Hopfinger et al.[4] The current commercial version

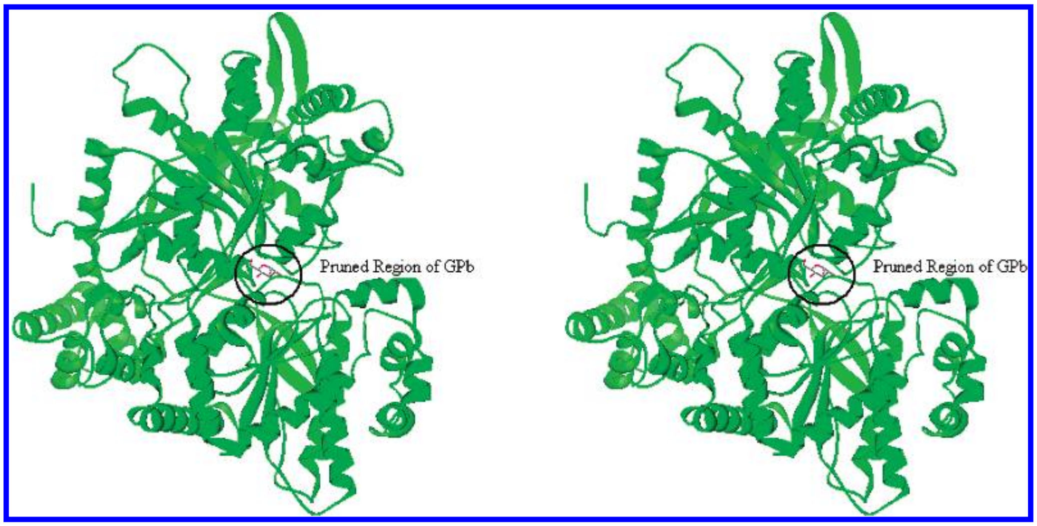


Figure 1. A graphical representation of the glucose bound GPb complex. The (schematic) protein is in green. The bound glucose is colored by elements: oxygen atoms are in red and carbon atoms in gray. The pruned receptor model of the protein is defined by the structure within the black circle.

of the 4D-QSAR software package,[17] version 3.0, was used to perform many RD and RI steps in the study. When the geometry of the receptor is taken into consideration (the RD-4D-QSAR formalism), modifications are necessary to the RI-4D-QSAR methodology. One key addition in extending the RI-4D-QSAR strategy to the RD-4D-QSAR methodology is applying the receptor geometry pruning technique that is an adopted and modified approach to the one proposed by Tokarski and Hopfinger.[3]

Step 1: Receptor Pruning and Atom Charge Assignment. The complete structure of the glycogen phosphorylase-b glucose bound complex consists of 13598 atoms including hydrogens. It is extremely computationally intense to sample ligand−receptor geometries using molecular dynamic simulation (MDS), or any other sampling techniques, on a molecular system of this size. However, such sampling is necessary to gain a meaningful thermodynamic averaged ensemble profile of such a system which is a major component to the “fourth” dimension in the 4D-QSAR paradigm. Fortunately, the ligand−receptor interactions are relatively short range as compared to the size of the entire receptor, and, thus, largely only involve the “lining” of the ligand binding site. Receptor pruning is an approach for achieving reasonable ensemble profiling and performing practical RD-4D-QSAR analysis in terms of time and computational resources. Basically, pruning is a preprocessing operation to scale down the protein to a manageable size structure containing the lining of the binding site before undertaking the actual 4D-QSAR phase of the study.

Receptor pruning is performed using HyperChem 5.01. Atom 1 of the bound glucose inhibitor is chosen as the center of the pruning volume, see Figure 1. It has been demonstrated in a previous study that the inhibitor-GPb interaction energies show no significant change after the size (radius) of the pruned GPb receptor increases beyond 10 Å.[2] Hence, residues more than 12 Å away from this center, are “cut off” subject to any GPb residues having at least one atom in the 12 Å region being completely included in the pruned receptor (the binding lining) model. The 2 Å “safety” factor (12 Å−10 Å) is employed as a “frozen shell” and is discussed below. To retain the integrity of the local geometric environment of the receptor, residue fragments separated by less than five intervening residues are connected by the missing residues. The important 280 loop of GPb (residues 282−286) is retained by including the residues cut off in the pruning process. Zero mass and zero partial charge hydrogen atoms are used to complete the open ends of the residue fragments of the pruned model.

The final pruned receptor model of GPb is composed of 1999 atoms from 124 residues. All atoms, except the fictitious hydrogen atoms added at the residue fragment ends of the pruned receptor model, are assigned AMBER partial charges.[18] Ionizable residues are neutralized to compensate for counterions complexing about the protein. All mobile counterions are considered to be located near the ionizable groups of the protein in this study. Thus, the protein is effectively neutral. The reduction of charge for an ionizable residue is distributed over the atoms of the actual ionizable group of the residue. The GPb cofactor, pyridoxaldehyde phosphate (PLP), is also retained in the receptor pruning process. The largest inhibitor of the training set, compound 43 in Table 1, is shown docked into the pruned inhibitor binding site in Figure 2, to demonstrate that training set inhibitors fit into and satisfy the 12 Å cutoff for the pruned receptor model.

Step 2: 3D Model Building of the Training Set and Test Set Inhibitors. The 3D structure of the bound glucose inhibitor was extracted from the crystal structure of the complex (2GPB). According to the crystal structure investigations,[14] most of the training set compounds, for which the crystal complex structures have been determined, have very similar bound conformations and bind (align) to the receptor in a highly analogous manner to glucose. The geometry of glucose bound in the catalytic site of GPb was, therefore, used as the template for the starting conformation and binding alignment for each of the training and test set inhibitors. That is, the other training and test inhibitors were built from the extracted glucose inhibitor by simply adding the corresponding substituents on the predefined glucose scaffold using HyperChem 5.01.[19] Partial atomic charges of the inhibitors were computed using the semiemperical AM1 method.[20] A minumum energy conformation of each inhibi-

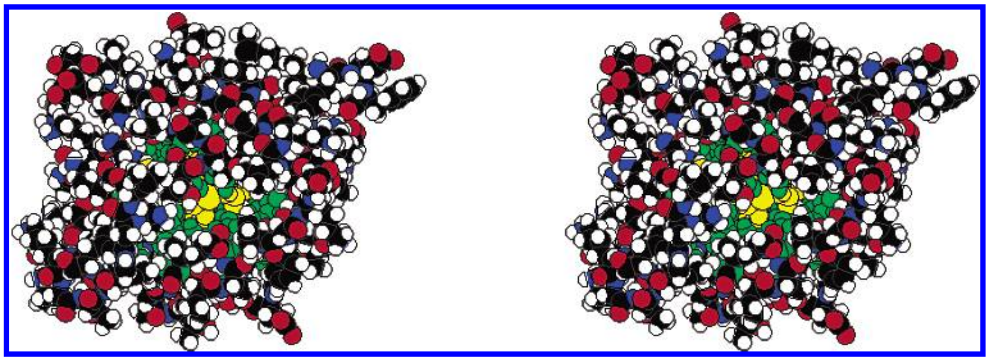


Figure 2. Complex of the pruned receptor model with inhibitor 43 bound. It is seen that inhibitor 43 (in gold) is buried inside the pruned receptor. The atoms in the green region of the pruned receptor have full flexibility which include atoms within 10 Å of atom 1 of inhibitor 43, the cofactor PLP and loop 280. The atoms outside the green region are constrained and form a rigid outer shell to the pruned receptor model.

tor, before it is docked into the binding site of GPb, was determined using the MM steepest descent method[20] and by the bound conformation as the minimization starting point.

Step 3: Ligand Docking. All the inhibitors of the training set are glucose derivatives which make it possible to align them to the binding topology of glucose to generate an initial binding geometry of each inhibitor-GPb complex. The inhibitors were aligned to the invariant three atoms (6, 2, 4) of the bound glucose, see the structures listed in Table 1. The bound glucose was then removed from the binding site, and the corresponding set of inhibitor-bound complexes were generated.

Step 4: Atom Type (Interaction Pharmacophore Element, IPE) Assignment. All the atoms of the pruned GPb-inhibitor complex are assigned interaction pharmacophore elements (IPEs), which are defined as follows: (a) any type of atom (*any*); (b) nonpolar atom, (*np*); (c) polar atom of positive partial charge, (*p*+); (d) polar atom of negative partial charge, (*p*−); (e) hydrogen bond acceptor, (*hba*); (f) hydrogen bond donor, (*hbd*); (g) aromatic atoms, (*ar*). The atom type (IPE) assignment permits the classification of enzyme−inhibitor interactions.

Step 5: Constraint of Selective Receptor and Inhibitor Atoms. The pruned receptor model of GPb is composed of 15 residue fragments between which the protein binding topology is lost in the pruning process. Overall, the residues located at the ends of the fragments, and those also located at the outer edges of the pruned receptor model, have considerably more conformational flexibility than when they are part of the complete parent protein. To facilitate the pruned receptor model retaining the highest conformational similarity to the complete protein, a constraining approach called *coordinate fixation* has been applied. Another constraining method called *fictitious mass assignment*, which assigns every atom of the pruned receptor model an identified heavy mass, was also applied and gave satisfactory results in an earlier study.[3] However, the *fictitious mass assignment* method tends to overconstrain residue side-chains and, in particular, suppress the lining of the binding site's ability to reorientate and reshape itself to better accommodate a ligand. In other words, insufficient flexibility is given to the side chain atoms (particularly those of the binding site lining) during the conformational ensemble sampling.

In the *coordinate fixation* method, the coordinates of only those atoms which are outside a 10 Å inner sphere of the 12 Å radius spherical pruned ligand−receptor complex are fixed, as is shown in Figure 2. These rigid outer layer atoms prevent the whole model complex from undergoing large/or meaningless motions and adopting unrealistic geometries. However, the remaining inner atoms of the pruned ligand−receptor complex are not excessively restricted from generating reasonable conformational changes throughout the complex. All of the atoms inside the "rigid outer layer" retain their intrinsic conformational properties. The translational and rotational movements of the bound inhibitor in response to the pruned GPb model are constrained at atoms (6, 2, 4) of the glucose ring, see top of Table 1. This binding alignment constraint is necessary to permit the generation of consistent grid cell occupancy descriptors (GCODs) across the training set.

Step 6: Molecular Dynamic Simulation (MDS). MDS is used to generate the conformational ensemble profile (CEP) of each pruned ligand−receptor complex. Before performing the MDS in the study, geometry optimization was performed using both MM steepest descent and conjugate gradient methods to obtain the lowest potential energy for each individual complex. Each resulting low-energy complex was used as the initial structure in the corresponding MDS. Both geometry optimization and MDS were done using the MOLSIM package.[21] An MDS of 20 ps sampling time with time-step intervals of 0.001 ps was performed for a total sampling of 20 000 conformations of each pruned GPb-inhibitor complex. A simulation temperature of 300 K and a molecular dielectric of 3.5 were used in the MDS. The atomic coordinates of each complex conformation and its energy were recorded every 0.01 ps for a total of 2000 frames for constructing the CEP and corresponding GCOD trial descriptor set of each model ligand−receptor complex.

Step 7: Alignment in a Binding Site. In the general RI-4D-QSAR method, alignment freedom is addressed by treating it as a search and sample operation analogous to conformational profiling. To date only an ordered three-atom alignment rule has been used, but more than one alignment can be readily made. The RI-4D-QSAR methodology has the ability to rapidly evaluate all proposed alignments in term of the relative significance of the corresponding final RI-4D-QSAR models generated as part of the QSAR analysis. However, in this RD-4D-QSAR study, the lining of the binding site is included, and inappropriate test alignments can introduce sterically forbidden overlaps of parts of an inhibitor with parts of the binding site model. Overall,

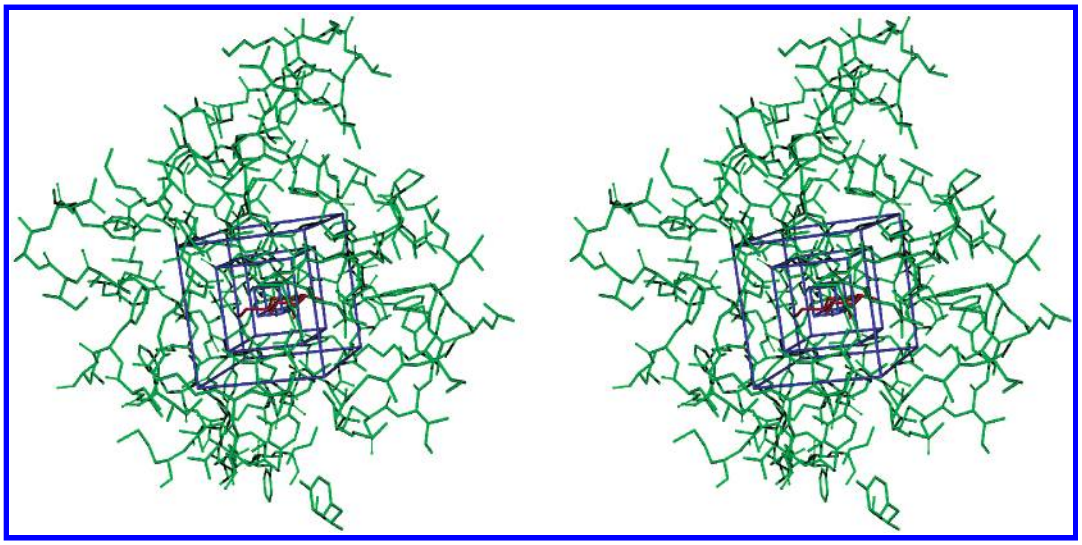


Figure 3. A conceptual representation of the local lattice generation process (step 9). The cubes, local grid lattices with side lengths of 2, 6, and 10 Å are embedded in the pruned receptor model.

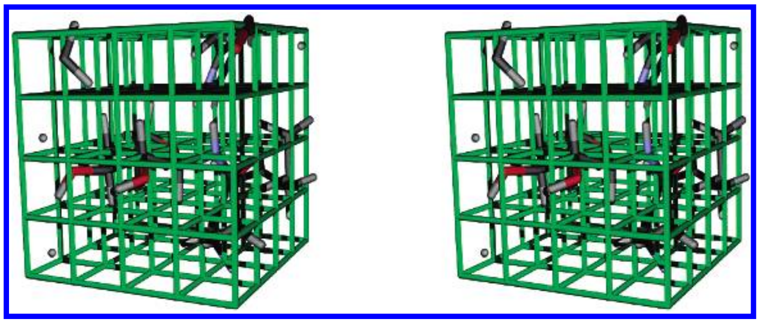


Figure 4. Graphical representation of a grid cell analysis and corresponding GCOD descriptors generation using local lattices (step 9), see Figure 3. Each grid cell of the lattice has a side length of 1 Å. Only one conformation of the GPb-inhibitor complex is shown in order to give a conceptual view of the process. In the RD-4D-QSAR paradigm each conformation in the CEP is placed in the grid cell lattice, and each final GCOD value is the normalized occupancy value from all CEP conformations.

flexible and/or nonequivalent atoms among inhibitors are not good alignment atoms because of the high probability of introducing structural damage to the pruned receptor−ligand complex. Moreover, it is known from a previous study[14] that the glucose ring of most inhibitors in Table 1 bind to the receptor in a very similar manner. This constraint applied to the GPb-inhibitor complexes permits only one unique ordered three-atom alignment no matter which three "constant" atoms are selected. Atoms (6, 2, 4) of glucose ring of the inhibitors, see Table 1, is the only allowed alignment of this RD-4D-QSAR study.

Step 8: Grid Analysis. Each CEP conformation of a pruned receptor−ligand complex is placed in the reference grid cell lattice according to the alignment under consideration in step 7. Again, in this study, there is only one alignment, namely atoms (6, 2, 4) of the glucose ring. The resolution of the grid cell lattice is 1 Å. The normalized absolute occupancy of each grid cell by each IPE atom type over the CEP for the alignment provides the trial pool of RD-4D-QSAR independent variables referred to as the grid cell occupancy descriptors, GCODs.

Step 9: Trial Descriptor Pool Generation. The grid cell occupancy profile from step 8 contains 60 312 GCODs. This is an excessive number of GCODs from which to readily extract the significant descriptor set to build a RD-4D-QSAR model. Moreover, the likely possibility of highly correlated GCODs due to ligand−receptor interactions, induced-fit ligand and/or receptor conformational changes, and ligand modulated receptor allosteric effects can lead to a large number of QSAR models with similar measures of significance which further complicates both model validation and interpretation.

The previous RI-4D-QSAR analysis of the training set in Table 1 indicates that the most important GCODs are located very close to the glucose ring.[5] Thus, initially focusing this RD-4D-QSAR study in small regions (local lattices) centered at the binding (GPb catalytic) site, instead of the entire ligand−receptor complex, can provide an opportunity to explore such a "local" binding phenomenon. Generation of a local lattice is similar to the receptor pruning process. Atom 1 of the glucose ring (see Table 1) is chosen as the center of the local lattices with composite side lengths of n Å ($n = 2$, 4, 6, 8, 10, 12) "cut out" as cubes from the complete reference grid lattice, see Figure 3. Structural information contained within each of these cubes is then defined over its local lattice and used to generate a trial pool of GCODs for constructing corresponding RD-4D-QSAR models, see Figure 4. All of these cubes are ranked against one another in term of the relative significance of the corresponding RD-4D-QSAR models.

Step 10: Partial Least-Squares (PLS) Regression Analysis. PLS[22] is employed as a data reduction tool to identify the most highly weighted GCODs from the entire set generated for a given local lattice (cube). The use of PLS in RD-4D-QSAR is identical to what is done in a RI-4D-QSAR analysis but repeated for each cube.

Step 11: Construction of RD-4D-QSAR Models/GFA-MLR-GFA. The N most highly ranked PLS GCOD descriptors determined in step 10 for a given cube are chosen to form the trial descriptor basis set for application of the genetic function approximation (GFA) model optimization.[23] If the number of GCODs in the grid cell occupancy profile for a given cube is less than 800, all those GCODs are retained for the GFA descriptor pool. Otherwise, the top 800 GCODs from the PLS analysis (step 10) are used. Diagnostic measures to analyze the resultant QSAR models are determined as part of the GFA optimization. The diagnostic measures include descriptor usage as a function of crossover operation, linear cross-correlation among descriptors and/or biological activity measures, and measures of model significance including the correlation coefficient, r^2, leave-one-out cross-validation correlation coefficient, q^2, and Friedman's lack of Fit, LOF.[24] The GFA searches for combinations of descriptors (GCODs) which score well, rather than identifying good individual high-scoring descriptors. Thus, it is possible that one, or more, of the descriptors in a good GFA RD-4D-QSAR model does not individually contribute significantly to the overall quality of a good GFA model. These types of low-scoring descriptors can cause confusion when trying to interpret a model.

In this study, in addition to building a RD-4D-QSAR model with high predictive power, identifying the spatial pharmacophores and corresponding ligand−receptor interactions are also high interest objectives. To facilitate realizing these objectives, the top 10 models from GFA optimization, which have the highest q^2, without data overfitting, are determined and investigated for each cube.

The number of times each unique GCOD is used in the top ten RD-4D-QSAR models is recorded. The number of top models investigated can be varied. The selection of 10 seems to capture the major GCODs (3D pharmacophore sites). Any GCOD used more than once for each cube among the top 10 models is retained for the next model building operation employing multiple linear regression analysis (MLR). Since the GCOD are determined by GFA, all of them may be significant. Thus, an appropriate MLR strategy, backward elimination,[25] is performed to filter out the relatively less important GCOD descriptors from the composite set found by application of GFA optimization to each cube. The selection criterion in the backward elimination step is the F-test. In the well-established MLR backward elimination strategy, a full model is built using all available descriptors. Then every descriptor is eliminated, one at a time, to build a series of one-term-less models. The one-term-less model with the least loss in significance, i.e., smallest F value, or largest p value, is then chosen as the new "full" model for next elimination step. The elimination process continues until removing any of the remaining descriptors in a model results in a loss of significance exceeding a threshold, known as the predefined significance level. The threshold was set as $p < 0.05$ in this study, where p denotes the confidence level.

The MLR backward elimination procedure was performed using SAS 8.1.[26] This technique is particularly useful in uncovering spurious relationships between the independent and dependent variables of a training set. However, one of the major drawbacks of the backward elimination process is that it tends to generate overfit models. Thus, in an effort to obtain nonoverfit models, GFA is subsequently performed a second time using all GCODs in the MLR model after backward elimination as a descriptor pool. The GFA model with a high q^2 and appropriate number of GCOD, defined as the model size for which additional GCOD does not increase the q^2 of a model, is selected as the best overall model. This multiple-step receptor-dependent model optimization process is referred to as GFA-MLR-GFA model building.

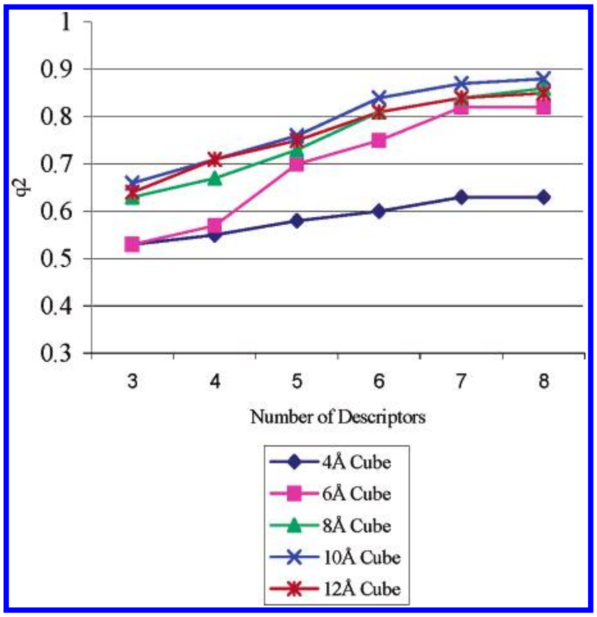


Figure 5. Plot of q^2 vs the number of descriptors in the corresponding RD-4D-QSAR models. The models are obtained in the first GFA optimization process of the GFA-MLR-GFA processing scheme using trial descriptor pools generated from the local lattices (cubes), see Figures 3 and 4, of varying sizes.

Step 12: Active Conformation Postulation. The final step in the 4D-QSAR paradigm is to hypothesize an "active conformation". In RD-4D-QSAR analysis, the active conformation of the composite ligand−receptor complex is postulated. This is achieved by first identifying all conformer states sampled for each complex that are within ΔE of the global minimum energy conformation of CEP. The ΔE was set as 2 kcal/mol in this study. Then the single complex conformation within ΔE that predicts the highest activity (inhibition potency) for a particular RD-4D-QSAR model is chosen as the active conformation of the complex with respect to that model.

RESULTS

The quality of the best RD-4D-QSAR models, represented by q^2, generated from the initial GFA optimization, using GCOD descriptor pools generated from local lattices represented by cubes of 4 Å through 12 Å, see Figures 3 and 4, are presented in Figure 5. When additional GCOD terms are added to a model, r^2 always increases, but q^2 may stop increasing, or may even decrease. The critical model size where q^2 stops increasing signals possible model overfitting and defines the number of descriptors that should be included in a good predictive model. The q^2 versus number of descriptors plot, Figure 5, for the first GFA procedure in

Table 3. Summary of the Results for Each Step of the GFA-MLR-GFA Process in the Determination of the Size of the Optimal Local Lattice (Cube)

results for each step of the GFA-MLR-GFA	6 Å cube	8 Å cube	10 Å cube	12 Å cube
number of unique GCODs after first GFA	11	16	18	20
number of GCODs survived MLR ($p < 0.05$)	8	11	11	12
quality of final model (number of terms)	7	6	6	6
quality of the final model (r^2)	0.88	0.85	0.88	0.86
quality of the final model (q^2)	0.82	0.81	0.82	0.82

Table 4. Linear Cross-Correlation Matrix of the GCODs in Eq 2

	GC1(0,−3,2,any)	GC2(1,−2,2,hbd)	GC3(−3,−1,3,p+)	GC4(−1,−2,−3,any)	GC5(1,4,3,any)	GC6(1,4,4,p+)
GC1(0,−3,2,any)	1.00					
GC2(1,−2,2,hbd)	0.28	1.00				
GC3(−3,−1,3,p+)	−0.24	−0.12	1.00			
GC4(−1,−2,−3,any)	−0.06	−0.05	−0.19	1.00		
GC5(1,4,3,any)	0.14	0.01	−0.08	0.18	1.00	
GC6(1,4,4,p+)	−0.17	−0.17	0.01	0.07	−0.29	1.00

the GFA-MLR-GFA model building process shows optimum model sizes for all local lattices (cubes). The descriptor pool generated from the 2 Å cube contains only 18 GCODs, and the best corresponding GFA models, which are not plotted in Figure 5, are extremely poor having q^2 values less than 0.1. Notice in Figure 5 that the curves for the 6 Å, 8 Å, 10 Å, and 12 Å cubes are very near one another, while the plot for the 4 Å cube is distant from them. Models obtained from the GCODs of the 4 Å cube have low q^2 values reaching only 0.6. This finding indicates that not all the inhibitor-GPb interactions involve structures located in a region smaller than that of the 4 Å cube. There is a clear indication that increasing the size of the local lattice (cube) may lead to more significant RD-4D-QSAR models. Thus, investigations of larger local lattices have been made, and the models tested for overfitting. For these optimized models 7, 6, 6, and 6 descriptors have been obtained from 6 Å, 8 Å, 10 Å, and 12 Å cubes, respectively. GCODs, from all these cubes, whose usage among the top 10 models occurs more than once, have been pooled together to generate the trial descriptor set for the backward elimination MLR analysis.

The results from each step of the three-step GFA-MLR-GFA model construction process are summarized in Table 3. After the first GFA application, models generated from the 6 Å cube must have one more term (7 terms) than the models from the other cubes in order to gain comparable q^2 values to the larger cube models. This finding suggests the 6 Å cube may still be too small to contain an optimum structural description, as represented by GCODs, of the GPb-inhibitor binding process. The final model of the three-step model construction procedure demonstrates the binding information of the 6 Å cube is insufficient. The best model generated from the 6 Å cube has an r^2 of 0.88 and a q^2 of 0.83 for 7 GCOD terms. But models from larger cubes having similar statistical significance are smaller in size, making the model from the 6 Å cube less attractive to further explore. However, it is not necessary to select the best model from the largest cube. The final best models from cubes of size 8 Å, 10 Å, and 12 Å are all comparable in quality to one another. Moreover, the number of GCODs surviving the backward elimination step is all about the same for each of these cubes, even though the 10 Å cube is twice the volume of the 8 Å cube, and the 12 Å cube is about three times larger. Presumably no additional inhibition information can be extracted, using GCODs, for local lattices (cubes) larger than 8 Å cube.

Overall, a cube of size 8 Å is selected as the most appropriate local lattice to generate the GCODs pool for model construction. In fact, the best model has already been determined as part of the process to select the "correct" cube. The optimum RD-4D-QSAR model is

$$\begin{aligned}\Delta G = {} & 2.4\mathrm{GC1}(0, -3, 2, \mathrm{any}) + 6.4\mathrm{GC2}(1, -2, 2, \mathrm{hbd}) + 8.7\mathrm{GC3}(-3, -1, 3, \mathrm{p+}) + \\ & 2.5\mathrm{GC4}(-1, -2, -3, \mathrm{any}) + 2.9\mathrm{GC5}(1, 4, 3, \mathrm{any}) - 1.8\mathrm{GC6}(1, 4, 4, \mathrm{p+}) + 3.03 \\ & n = 47,\ r^2 = 0.85,\ q^2 = 0.82 \qquad (2)\end{aligned}$$

In eq 2, ΔG is the free energy of GPb-inhibitor binding. The GC*i* (x, y, z, X) is the *i*th GCOD with coordinates (x, y, z) and atom type X as defined in step 4. The GCODs define the spatial locations of the key sites that increase/decrease the free energy of binding if occupied by the correct types of atoms. As such, the six GCODs define the spatial pharmacophore of the training set which includes both ligand and receptor contributions.

Chance correlation is one of the biggest problems when the trial descriptor pool is excessively large. To demonstrate the reliability of the RD models, data scrambling is performed.[6] The models obtained after the ΔG values are randomly reassigned among the training set inhibitors are of less significance with an average q^2 of 0.5 and a maximum of 0.6, which corresponds to a 40% decrease compared to that of the best RD-4D-QSAR model (eq 2). Therefore, a valid correlation between the GCODs and the ΔG values of the training set has been captured by the RD model given by eq 2.

The cross correlation matrix of the GCODs of the RD-4D-QSAR model given by eq 2 is presented in Table 4. No significant collinearity between any pair of GCODs is found. Thus, each GCOD of eq 2 contains unique ligand−receptor binding information.

The predicted ΔG of each inhibitor-GPb complex of the training set is calculated using eq 2. The difference between the predicted and observed ΔG values, the residual of fit, are plotted in Figure 6. The RD-4D-QSAR model predicts well for the training set compounds.

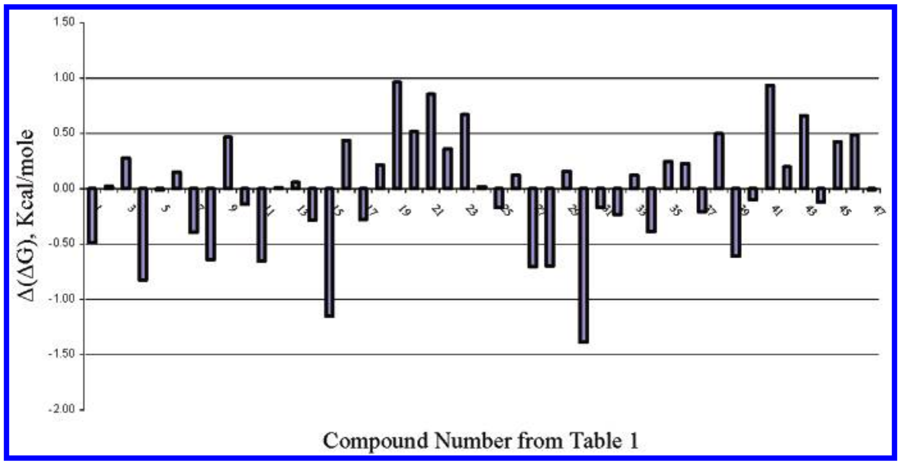


Figure 6. The residuals of fit plot for the training set using the best RD-4D-QSAR model given by eq 2.

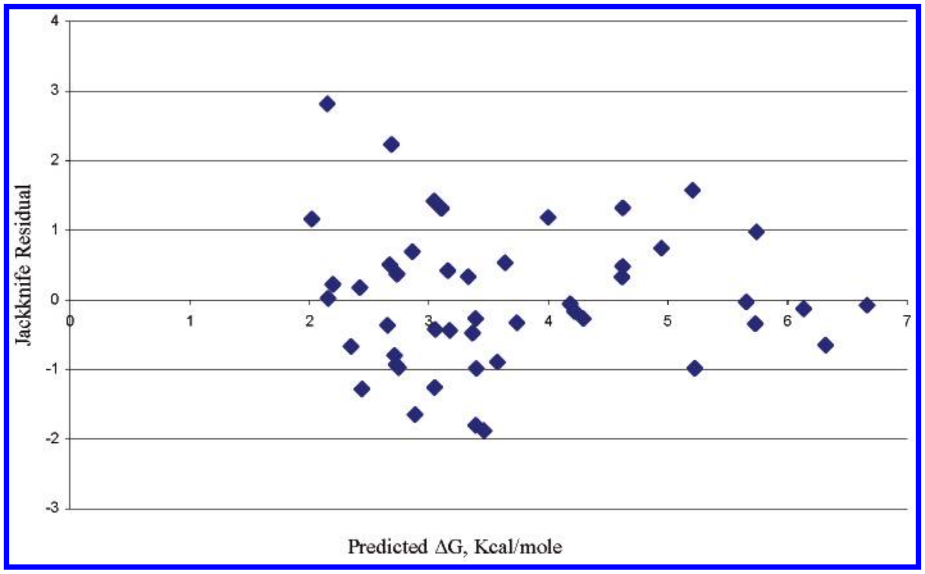


Figure 7. The jackknife residuals of fit plot of the training set using the best RD-4D-QSAR model, eq 2.

The Jackknife residuals[25] have also been calculated and are plotted against the predicted activity, Δ*G*, in Figure 7. This procedure is performed in this study to test the assumption of linearity in the relationship between the GCOD descriptors and Δ*G*. No systematic trends of any kind are observed in the plot in Figure 7. This finding supports the hypothesis that employing a linear model to correlate the GCOD descriptors to Δ*G* is appropriate. Another application of the Jackknife plot is to identify outliers. Data points in the Jackknife plot that lie far from the rest of the data should be flagged and investigated further in order to find the effects of these "abnormal" data on a predictive model. All of the data points in Figure 7 are approximately evenly distributed on both side of the predicted activity axis (Jackknife residual = 0) and there are no "abnormal" data point.

The active conformation of each complex formed by each training set inhibitor bound to the pruned GPb receptor model has been postulated using the best RD-4D-QSAR model, namely, eq 2. Graphical representations of the active conformations of the complexes for training set inhibitors 24, 1, 10, and 39 of Table 1 are shown in Figures 8 and 9, together with the 3D pharmacophore representation of the RD-4D-QSAR model, eq 2. These four inhibitors comprise two of the most, one average and one of the least potent inhibitors. In addition to spanning the range of observed inhibition, a more important reason for choosing to focus on the complexes of these four inhibitors is that, in composite, they possess high GCOD values with respect to eq 2. Hence, these inhibitors offer a good perspective regarding how the GCODs of eq 2 identify the key interacting chemical groups in individual inhibitor-GPb binding processes. To be specific, complex 24 has high GC1, GC2, and GC4 values; complex 1 significantly occupies GC5, while complex 10 occupies GC3 and complex 39 GC6. Complex 5 is also seen in Figure 9 to demonstrate the important general feature of a unique induced-fit between each specific ligand and the receptor, which GC1(0, −3, 2, 0) particularly captures for complex 5.

The pruned receptor with the most active inhibitor, compound 24, and the six GCOD descriptors in the best RD-4D-QSAR model, eq 2, are presented in Figure 8. Blue cubes represent GCODs with positive regression coefficients, i.e. activity enhancing GCODs, while the red cube represents a GCOD with a negative regression coefficient, i.e., an activity decreasing GCOD. It is seen that three of the GCODs are occupied by the inhibitor and the other three GCODs by the pruned receptor. In Figure 9 only the directly interacting residues of the pruned GPb receptor binding site are shown with each inhibitor in order to give a straightforward view of the key bindng sites in each of the complexes.

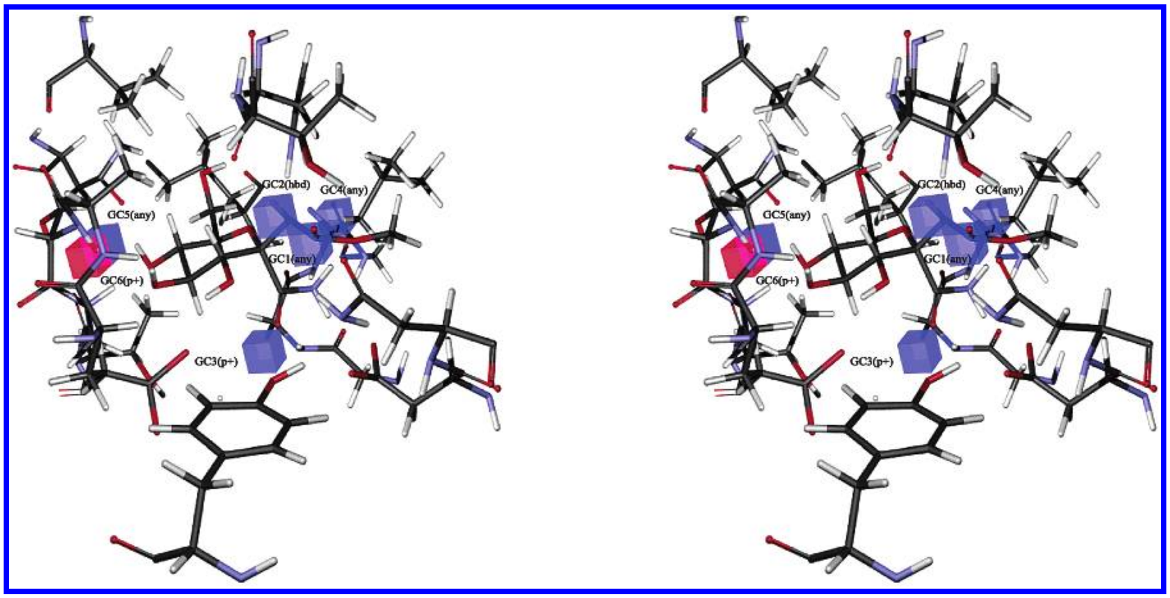


Figure 8. A visual representation of the spatial relationship between the binding site of complex 24 (the pruned GPb receptor model with the most active inhibitor in the training set, compound 24) and the GCODs of the best RD-4D-QSAR model, eq 2. GCODs having positive regression coefficients in the model are blue, while a GCOD having negative regression coefficient is red.

Perhaps the most interesting GCOD in Figure 9 is GC1(0, −3, 2, any). The occupancy value of a GCOD with IPE type "any" is the sum of the occupancy values of all unique IPE GCODs at the same position. An inspection of the occupancy profile at position (0, −3, 2), given in Table 5, reveals there is no clear pattern for the composition of (0, −3, 2, any). However, it is seen that GC1(0, −3, 2, any) equals the sum of (0, −3, 2, np) and either (0, −3, 2, hba) or (0, −3, 2, p-) for complex 24. In Figure 9A, GC1 is situated at the −CO− group of compound 24, which matches the IPE types of its occupancy profile: the oxygen atom contributes to the occupancy of (0, −3, 2, hba) or (0, −3, 2, p-), while the carbon atom contributes to the occupancy of (0,−3, 2, np). GC1 is located only 2.4 Å from an NH_2 hydrogen atom of the side chain of Asn 284, suggesting the formation of a hydrogen bond between the NH_2 of Asn 284 and the carbonyl oxygen of compound 24.

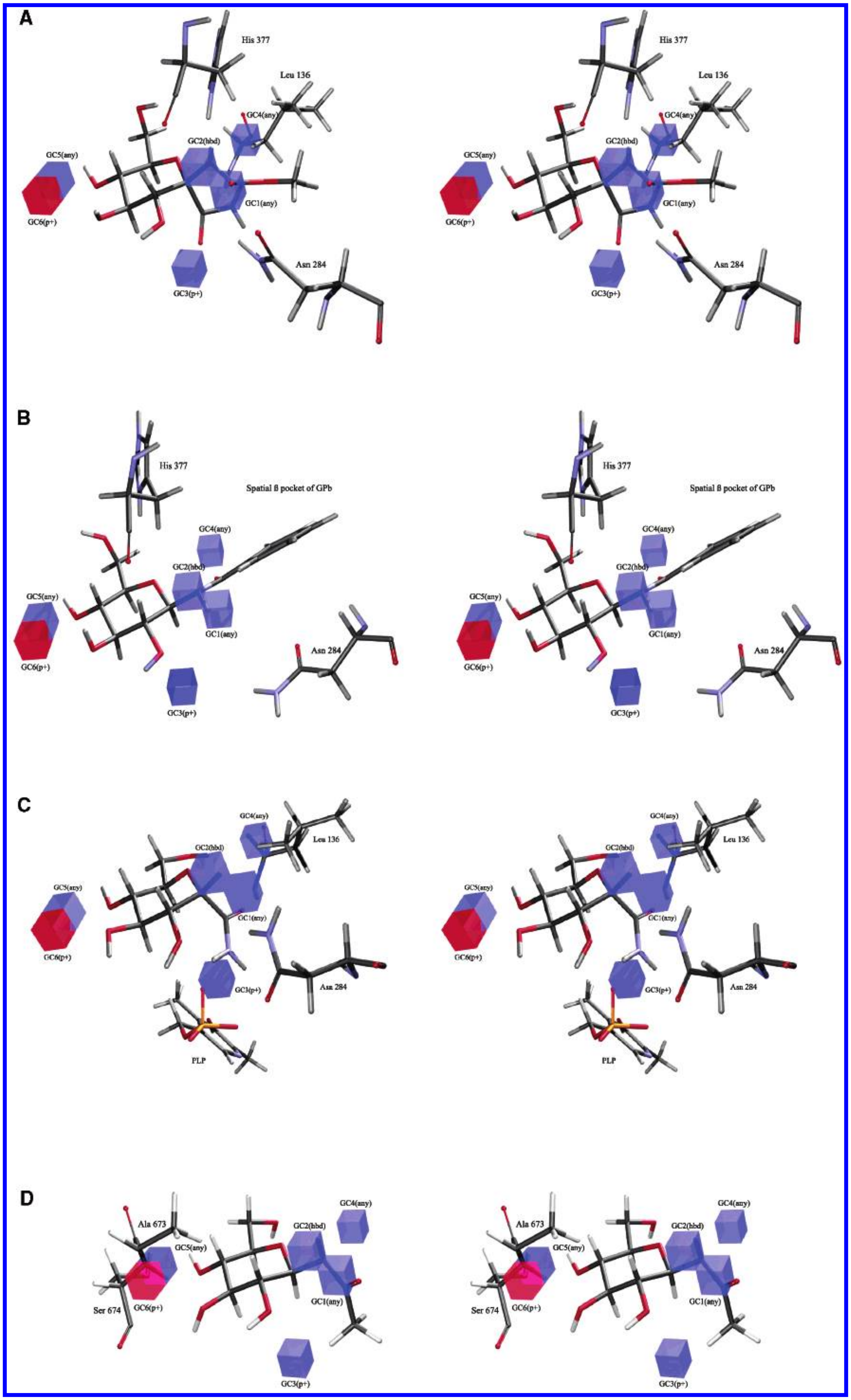
In contrast to compound 24, the value of GC1 for complex 5 arises exclusively from GC(0, −3, 2, hbd) or (0, −3, 2, p+). Figure 9B illustrates that the orientation of Asn 284 in complex 5 has changed from that of complex 24. The −CO− group of the Asn 284 side chain moves 0.5 Å closer to the inhibitor, while the side chain $-NH_2$ group moves 1.3 Å away from the inhibition site. These reorientations result in a better alignment for forming a hydrogen bond between the oxygen of the Asn 284 side chain and the proton of the −NHCO− of compound 5. Thus, the loss of the strong hydrogen bond between the $-NH_2$ of Asn 284 and the −CO− of the inhibitor, as seen in complex 24, is partially compensated when avoiding steric hindrance between the phenyl group of compound 5 and the Asn 284 as the phenyl group locates in the β pocket of the catalytic site of the receptor.

Realignment and conformational changes in Asn 284 to realize highest inhibitor-GPb binding cannot be derived from receptor-independent QSAR methods. The changes in receptor geometry observed in the simulation modeling of ligand−receptor binding are indicative of induced fits between GPb and inhibitors. GC1 demonstrates the possibility that same GPb-inhibitor interaction (hydrogen bonding) occurs at the same position (at coordinate (0, −3, 2)) but with different interacting groups (the −CO− of inhibitor 24 and amine group of Asn 284 or between the −NH− of inhibitor 5 and the −CO− of Asn 284).

GC2(1, −2, 2, hbd) is situated, see Figure 9A, at the "top" of the −NH− group of inhibitor 24 and is within hydrogen bonding distance (2.3 Å) to the backbone carbonyl oxygen of His 377. This observation suggests that a second ligand−receptor hydrogen bond of compound 24 with the −CO− of the backbone of His 377 contributes to the binding free energy of this inhibitor.

It is readily seen in Figure 9A,B that the proton of the −NHCO− moiety of an inhibitor across the training set can be the occupant of either GC2 (in complex 24) as in Figure 9A, or GC1 (in complex 5) as in Figure 9B. This finding is evidence for the necessity of performing MDS on a flexible ligand−receptor model in order to obtain reliable information on the preferable binding interactions. When compound 24 binds to GPb, the −NHCO− moiety of compound 24 forms two hydrogen bonds with GPb, one with the backbone −CO− of His 377, and the other with side chain amine hydrogen of Asn 284. If the substituent group(s) "beyond" the −NHCO− moiety is too big, as is the case for complex 5 and illustrated in 9B, both the inhibitor and the receptor undergo geometric changes to reduce steric hindrance. The geometric changes for complex 5 results in the loss of both of the hydrogen bonds found in complex 24, but one new hydrogen bond forms between the −NH− of compound 5 and the side chain oxygen of Asn 284.

The role of GC3(−3, −1, 3, p+) on ligand−receptor binding is illustrated in Figure 9C, for complex 10. GC3 is close to the NH_2 group of the α substituent of compound 10. In most complexes, the lack of α substituents for most of the training set compounds leads to a low occupancy value of this GCOD. GC3 is 3.6 Å from the carbonyl group of the side chain of Asn 284. An attractive electrostatic interaction, or weak hydrogen bond, is suggested by this GCOD. In addition, this GCOD is very close to the negatively charged


A
His 377
Leu 136
GC4(any)
GC2(hbd)
GC5(any)
GC6(p+)
GC1(any)
GC3(p+)
Asn 284
B
His 377
Spatial ß pocket of GPb
GC4(any)
GC2(hbd)
GC5(any)
GC6(p+)
GC1(any)
Asn 284
GC3(p+)
C
GC4(any)
GC2(hbd)
Leu 136
GC5(any)
GC1(any)
GC6(p+)
GC3(p+)
Asn 284
PLP
D
GC4(any)
Ala 673
GC2(hbd)
GC5(any)
Ser 674
GC6(p+)
GC1(any)
GC3(p+)

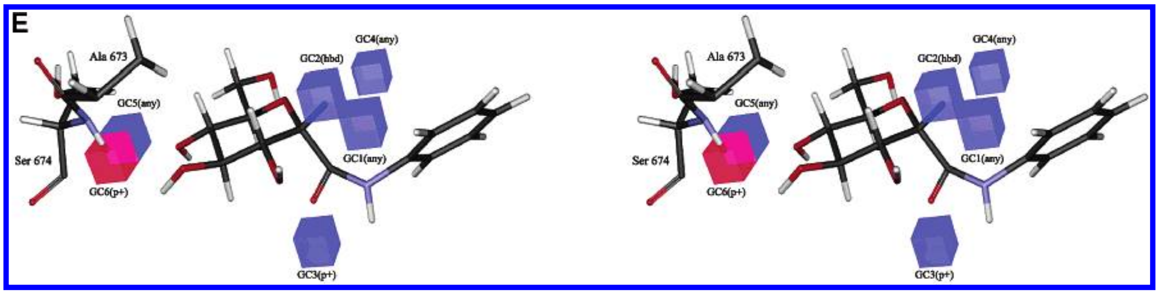


Figure 9. (A) A simplified visual representation of the spatial relationship between complex 24 and the GCODs of the best RD-4D-QSAR model. GC1(any) is situated on the −CO− of the β substituent of the inhibitor. GC2(hbd) is occupied by the −NH− of the same substituent. GC4(any) is close to Leu 136 of the receptor. (B) A simplified visual representation of the spatial relationship between complex 5 and the GCODs of the best RD-4D-QSAR model, eq 2. In this complex, GC1(any) is occupied by the −NH− of the β substituent of inhibitor 5. (C) A simplified visual representation of the spatial relationship between complex 10 and the GCODs of the best RD-4D-QSAR model. GC3(p+) is close to the amide group of the substituent of inhibitor 10. It is also in close proximity to Asn 284 and the cofactor PLP. Leu 136 is shown to demonstrate the difference in distances between this residue and GC4(any) from that found in complex 24, see part A. (D) A simplified visual representation of the spatial relationship between complex 1 and the GCODs of the best RD-4D-QSAR model. The backbone −NH− of Ser 674 occupies GC5(any). The β substituent of inhibitor 1 adopts a similar conformation to that of complex 24, see part A. The −CO− occupies GC1(any) and the −NH− occupies GC2(hbd). (E) A simplified visual representation of the spatial relationship between complex 39 and the GCODs of the best RD-4D-QSAR model. The proton of the backbone −NH− of Ser 674 occupies GC6(p+).

cofactor PLP. Electrostatic interaction between the phosphate group of PLP and the NH_2 proton may be captured by GC3 as well.

GC4(−1, −2, −3, any) is close to the backbone of Leu 136 as shown in Figure 9A. The location of this GCOD suggests that conformational and alignment changes of Leu 136 may be important in ligand−receptor binding. GC4, by itself, provides little specific information regarding what interaction may cause movement of Leu 136. However, by comparing parts A and C of Figure 9, it is seen that Leu 136 in Figure 9A is closer to the inhibitor than it is in Figure 9C. In turn, the hydrogen of the backbone −NH− of Leu 136 moves into a relatively close proximity (3.6 Å) to the oxygen of the α substituent of compound 24 and a modest hydrogen bond can form. The formation of this hydrogen bond gives the inhibitor additional binding affinity to GPb.

GCODs associated with a “constant” structure across the training set can be present in a RD-4D-QSAR model. In this study, GC5(1, 4, 3, any) and GC6(1, 4, 4, p+) are present at a region of common chemical structure in the receptor catalytic binding site, see Figure 9D,E. Due to a low cross-correlation coefficient of −0.29, these two GCODs are classified as independent variables even though they are only 1 Å from one another. Each GCOD contains unique structural information. Like GC4, these two GCODs are located in the receptor cavity away from the inhibitor, suggesting these two GCODs are monitoring conformational changes of the GPb receptor. GC5 and GC6 are located on the −NH− of the backbone of Ser 674 and are less than 3 Å from the hydroxy groups on carbons 3 and 4 of the glucose ring of most bound inhibitors. GC5 can be occupied by the backbone −NH− of Ser 674 and GC6 occupied by only the hydrogen atom of this −NH− group. The other interesting finding regarding these two GCODs is the opposite sign of their regression coefficients in the RD-4D-QSAR model, eq 2. The interaction between an inhibitor and the receptor lining of GPb seems to be influenced by a subtle interaction balance with inhibition activity increasing through hydrogen bonding but decreasing due to steric repulsion with Ser 674. Ser 674 adjusts its conformation and alignment to adopt a best binding orientation for the backbone −NH−. A small change in conformation and/or alignment can result in a large loss in inhibition potency.

A test set has been assembled consisting of four low to medium potency inhibitors (T1-T4), three relatively potent inhibitors (T5-T7), and one compound (T8) whose inhibition is beyond the upper activity range of the training set. This test set is given in Table 2. The predicted inhibition potency for the test set compounds are shown in Figure 10. The RD-4D-QSAR model, expressed by eq 2, predicts well for all inhibitors in Table 2. To further test the predictive strength of the RD-4D-QSAR model, a statistical measurement, called the shrinkage on cross-validation,[25] $S(q^2)$ has been calculated. The expression to calculate $S(q^2)$ is given in eq 3, where $r_{\text{training—set}}$(pre, obs) and $r_{\text{test—set}}$(pre, obs) denote the correlation coefficients between the observed and the predicted activities of the training and test sets, respectively, using the best RD-4D-QSAR model, eq 2.

$$S(q^2) = r^2_{\text{training—set}}(\text{pre, obs}) - r^2_{\text{test—set}}(\text{pre, obs}) \quad (3)$$

Since no structure−activity information from the test set compounds is included in QSAR model construction, it is expected that a QSAR model will better predict the training set than the test set. Nevertheless, a model with high predictive capacity should largely describe the structure−activity relationship inherent in a test set to an extent comparable to that realized for the training set. $S(q^2)$ is a quantitative measure of the predictive breadth of a QSAR model relative to a given training set and a given test set. In this study, $r^2_{\text{training—set}}$(pre, obs) is 0.85, and $r^2_{\text{test—set}}$(pre, obs) is 0.77. Thus, $S(q^2)$ is only 0.08, suggesting the RD-4D-QSAR model, eq 2, has good predictive breadth.

An optimized RI-4D-QSAR model was constructed by performing a receptor- independent analysis, using the RD-4D-QSAR three-ordered atom alignment (6, 2, 4), and is given in eq 4.

$$\begin{aligned}\Delta G = {} & 2.4\text{GC1}'(0, -3, 2, \text{hbd}) + 4.7\text{GC2}'(1, -2, 1, \text{p+}) + \\ & 18.7\text{GC3}'(-3, -1, 0, \text{p+})\ 8.3\text{GC4}'(-1, -3, 3, \text{p+}) + \\ & 11.8\text{GC5}'(2, -1, -1, \text{np}) - \\ & 4.1\text{GC6}'(-4, -3, 2, \text{np}) + 2.85 \\ & n = 47,\ r^2 = 0.85,\ q^2 = 0.82 \end{aligned} \quad (4)$$

Table 5. Occupancy Profile of Grid Cell GC(0, −3, 2) as a Function of Inhibitor

compd no.	GC(0,−3,2,any)	GC(0,−3,2,np)	GC(0,−3,2,p+)	GC(0,−3,2,p-)	GC(0,−3,2,hba)	GC(0,−3,2,hbd)
1	0.152	0.148	0	0.004	0.004	0
2	0	0	0	0	0	0
3	0.64	0.625	0.01	0.005	0.005	0.01
4	0.455	0.455	0	0	0	0
5	0.31	0	0.31	0	0	0.31
6	0.255	0.225	0.005	0.025	0.025	0.005
7	0.63	0	0.63	0	0	0.63
8	0.705	0	0	0.705	0.705	0
9	0.425	0.425	0	0	0	0
10	0	0	0	0	0	0
11	0.33	0	0	0.33	0.33	0
12	0.145	0	0.025	0.12	0.12	0.025
13	0.18	0.18	0	0	0	0
14	0	0	0	0	0	0
15	0.005	0	0.005	0	0	0.005
16	0.53	0	0	0.53	0.53	0
17	0.035	0	0.035	0	0	0.035
18	0.245	0.245	0	0	0	0
19	0.195	0.005	0.02	0.17	0.17	0.02
20	0	0	0	0	0	0
21	0.005	0	0.005	0	0	0.005
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	**0.48**	**0.47**	**0**	**0.01**	**0.01**	**0**
25	0	0	0	0	0	0
26	0	0	0	0	0	0
27	0.005	0	0	0.005	0.005	0
28	0	0	0	0	0	0
29	0.15	0	0	0.15	0.15	0
30	0.095	0	0	0.095	0.095	0
31	0	0	0	0	0	0
32	0.12	0.12	0	0	0	0
33	0.475	0	0	0.475	0.475	0
34	0.19	0.19	0	0	0	0
35	0	0	0	0	0	0
36	0.235	0.235	0	0	0	0
37	0.205	0.205	0	0	0	0
38	0.475	0	0	0.475	0.475	0
39	0.025	0.025	0	0	0	0
40	0	0	0	0	0	0
41	0.235	0.01	0.005	0.22	0	0.005
42	0	0	0	0	0	0
43	0.004	0	0	0.004	0	0
44	0	0	0	0	0	0
45	0	0	0	0	0	0
46	0	0	0	0	0	0
47	0	0	0	0	0	0

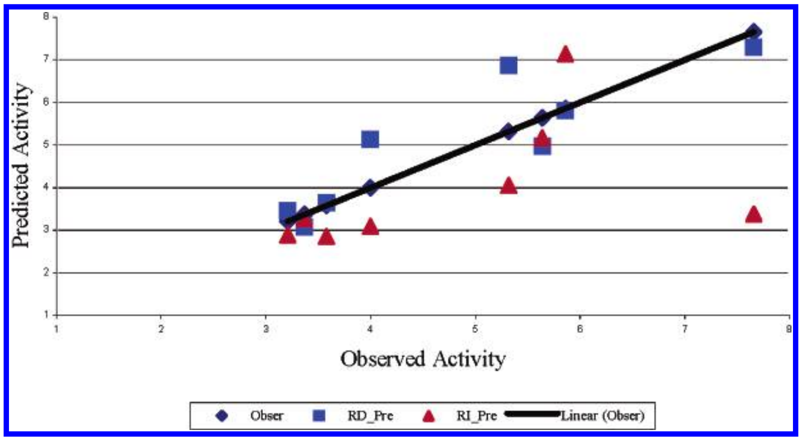


Figure 10. Predicted activity, Δ*G*, of the test set inhibitors using the best RD- and RI-4D-QSAR models. The Δ*G*'s of the test set compounds are ranked from low to high in the plot. The predicted Δ*G*'s using the best RI model are in red triangle. Predicted Δ*G*'s using the RD model are in blue squares.

Naturally, it is of interest to compare the RD- and RI-4D-QSAR models for the same alignment. One comparison between the RD- and RI-4D-QSAR models can be made by identifying how similar the pharmacophores (GCODs) of each model are to one another in space. The graphical representation of the RD and RI pharmacophores is shown

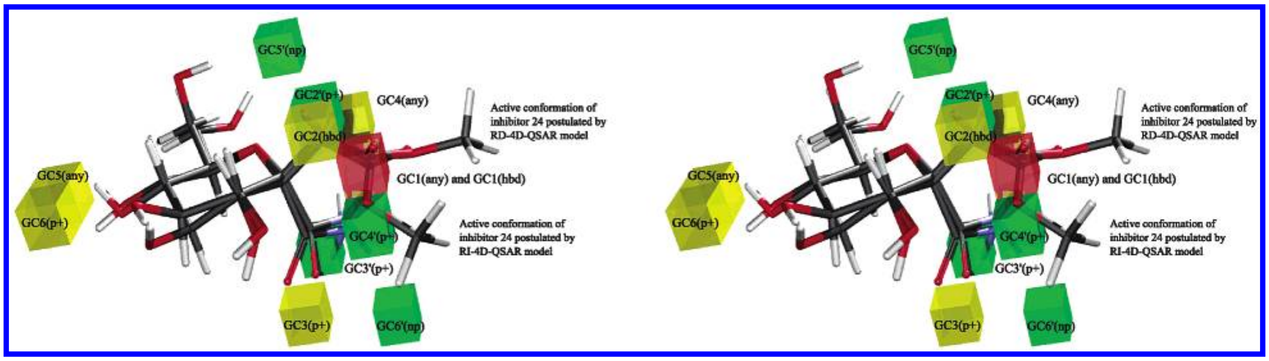

GC5'(np)
GC2'(p+)
GC4(any)
Active conformation of inhibitor 24 postulated by RD-4D-QSAR model
GC2(hbd)
GC5(any)
GC1(any) and GC1(hbd)
GC6(p+)
Active conformation of inhibitor 24 postulated by RI-4D-QSAR model
GC4'(p+)
GC3'(p+)
GC3(p+)
GC6'(np)
GC5'(np)
GC2'(p+)
GC4(any)
Active conformation of inhibitor 24 postulated by RD-4D-QSAR model
GC2(hbd)
GC5(any)
GC1(any) and GC1(hbd)
GC6(p+)
Active conformation of inhibitor 24 postulated by RI-4D-QSAR model
GC4'(p+)
GC3'(p+)
GC3(p+)
GC6'(np)

Figure 11. A GCOD comparison of the best RD- and RI-4D-QSAR models using the postulated active conformation of compound 24 from both models as references. RD GCODs are shown in gold and RI GCODs are in green. The cube in red indicates GC1 and GC1′ are superimposed.

in Figure 11. Only three GCODs of the RD-4D-QSAR model are occupied by atoms of the inhibitors. These three GCODs are used in the comparison to the six GCODs of the RI-4D-QSAR model given in eq 4. It is readily seen that GC1(0, −3, 2, any) and GC1′(0, −3, 2, hbd) share the same spatial location. The difference between these two GCOD descriptors is that GC1 has IPE type "any" while GC1′ has IPE type "hbd". The difference in IPE (atom) types likely occurs because of the larger differences in geometrical and IPE environments an inhibitor experiences in a RD-4D-QSAR analysis as compared to a RI-4D-QSAR study. GC1 contains binding information for the entire −NHCO− group of an inhibitor, while GC1′ only packages information for the hydrogen of the inhibitor −NHCO− group. GC1 in the RD-4D-QSAR model indicates that this group can act as either a hydrogen bond donor or acceptor in binding to GPb depending on the inhibitor. The RI model only captures part of hydrogen bonding binding features of the −NHCO− inhibitor group.

The close distance between GC2(1, −2, 2, hbd) and GC2′(1, −2, 1, p+) as well as the high similarity between their atom types, p+ and hbd, suggest these GCODs describe the same inhibitor-GPb interactions. However, these two GCOD descriptors are in slight disagreement as to where the interactions occur based on difference in their Z-coordinates. GC3(−3, −1, 3, p+) may be related to GC3′(−3, −1, 0, p+). Both of these GCODs are located in the same region of space, and both GCODs are a similar distance from Asn 284 and could also reflect an inhibitor proton having a favorable electrostatic interaction with the carboxyl group on the side chain of Asp 283. Despite the high similarities between three RD and RI GCOD pairs, it is important to remember that RI GCODs are not interchangeable with the RD GCODS due to the difference in structural information used in their determinations and their corresponding manner in "explaining" ligand−receptor interactions.

Another way to compare the RD- and RI-4D-QSAR models is to compare their respective residuals of prediction. The correlation coefficient of the residuals of prediction of the training set between the RD- and the RI-4D-QSAR models is 0.92, indicating a high similarity, and commonality, between these two models. A plausible speculative inference from this high correlation of residuals of prediction is that inclusion of the receptor geometry, in building a QSAR model, does *NOT* provide very much additional ligand−receptor binding information compared to that realized from the inhibitor training set structures by themselves. However, it must be remembered that this observation is made only with respect to the training set.

The test set predictions of the RD- and RI-4D-QSAR models are also shown in Figure 10. Both the RD- and RI-4D-QSAR models give about equal quality predictions over the five test set compounds (T1-T5) of medium inhibition potency which all have the same core structure, namely a glucose ring with flexible α and/or β substituents at atom 1, structure A (see Table 2), as found in all the training set compounds. However, when the core structures of test (new) compounds deviate, even in a minor manner from that of the training set, as is the case for test set compounds T6-T8 with structures B and C in Table 2, the corresponding accuracy of prediction varies markedly between the RD- and the RI-4D-QSAR models. Replacement of the two flexible α and β substituents on atom 1, structure A in Table 2, by structures B and C, results in less flexible analogue inhibitors relative to those of the training set. The inhibition potency of compound T8 of Table 2, which has a six-member pseudo spiro-ring, is overestimated by a $\Delta(\Delta G)$ 1.15kcal/mol by the RI-4D-QSAR model. If the inhibitor geometry is further constrained, the predictions of the RI-4D-QSAR model are even poorer. Compound 6 is underestimated by a $\Delta(\Delta G) =$ 1.27 kcal/mol, and the worst prediction for the test set is made on the most potent inhibitor, compound 7, its ΔG being underestimated by 4.25 kcal/mol using the RI-4D-QSAR model. However, the ΔG values for compounds 7 and 8 are both accurately predicted by the RD-4D-QSAR model being slightly underestimated by 0.06 kcal/mol (compound 7) and 0.41 (compound 8) kcal/mol. Thus, while the receptor geometry seems to provide little additional binding information for predicting inhibition potency of the training set compounds, as mentioned above, the receptor geometry appears to be crucial in screening compounds outside the training set with respect to both diversity in chemical structure and expanding the range of inhibition potency, especially to higher potency.

DISCUSSION

New techniques have been developed and applied in this investigation in order to successfully take the geometry of the receptor into consideration in building a RD-4D-QSAR model. The combination of protein pruning (step 1) and selective receptor atom constraints (step 5) retains a high approximation of the pruned receptor model to that of the entire 3D enzyme structure. That is, the pruned receptor model preserves the physiochemical environment of the "lining" of the inhibitor binding site. The root-mean-square, RMS, values for the atoms of the pruned GPb-inhibitor complexes before and after MDS range from 0.65 Å to 0.9 Å. This range of atomic displacements matches the motions of protein atoms under normal psychological conditions.[27]

The size of the pruned receptor can vary depending on the degree of flexibility of the specific receptor. Additional amino acid residues may be retained in the pruned model for a receptor with higher flexibility than one with less conformational freedom. One way to assess and validate a pruned receptor model is to determine if both the geometry and energy using the pruned receptor system are the same as computed using the complete receptor.[2]

An enormous amount of structural information (GCODs) is generated in a 4D-QSAR analysis due to the size of the receptor, the utilization of MDS profiling and the variety of IPE types. However, only information from the "functional" region, that is, the smallest region which provides definitive information about receptor-inhibitor binding, is essential. Investigating a series of local lattices (cubes) with various sizes (step 9), and centered about a bound inhibitor, is an effective strategy to identify this "functional" region. Using the structural information (GCODs) only from the functional region is more efficient, and computationally economical, than performing the RD-4D-QSAR analysis on the entire pruned receptor model. Moreover, the quality of the QSAR model can also be retained using only the functional region of the binding site. Overall, all important GCODs in the best RD-4D-QSAR model lie relatively close to the bound inhibitor, readily providing insights into the inhibitor-GPb binding interactions.

Model construction using the three-step GFA-MLR-GFA optimization procedure is an effective and insightful tool to sort through a large number of (GCOD) descriptors to find the most important ones and build corresponding good predictive models. In the GFA-MLR (back elimination)-GFA process, the first GFA analysis is applied as a filter to select a set of significant descriptors. These descriptors are determined by their usage in the top 10 models found in the first GFA step for each local lattice (cube). There is no clear rule how many models that are generated from the first GFA procedure should be considered for collecting an initial GCOD trial set for the MLR step. However, inferior models, having low q^2 values, are not suitable for further investigation because their descriptors have minimal significance. Therefore, an alternate strategy of descriptor selection for the second step might be to only use models within a small q^2 difference (less than 10% of the q^2 of the best initial GFA model) from the best model generated in the initial GFA optimization.

This relatively small set of descriptors from the first step is then used as a descriptor pool for MLR back-elimination for the ranking of descriptors in term of their relative significance. Backward elimination usually results in (slightly) overfit models which describe the structure–activity relationship well for the training set but not necessary for a test set. In this study, the MLR backward-elimination model generated from the 8 Å cube has 11 descriptors and a q^2 of 0.75 which is inferior to the best GFA model, eq 2, having only six descriptors and a q^2 of 0.81. Hence, a second GFA analysis is performed on the remaining descriptors of the MLR model in order to eliminate any overfitting introduced in the MLR step. The final MLR-GFA steps provide assurance that all GCODs in the final RD-4D-QSAR model are significant and can be used to direct inhibitor design which is the major application of structure-based design. Thus, the final model after the complete GFA-MLR-GFA model building process also meets the goals of a successful QSAR study: a good model with high predictive power having independent and significant descriptors possessing interpretative structural features.

A RI-4D-QSAR analysis builds QSAR models from the conformational ensemble profile (CEP) of the inhibitor training set which is explored in the absence of the receptor geometry. Still, RI-4D-QSAR models can significantly characterize the structure–activity relationship inherent to a training set. The best RI model, eq 4, has a q^2 value of 0.82 for six GCOD descriptors all of which are based solely upon inhibitor atom grid cell occupancy. In contrast, the best RD-4D-QSAR model, shown in eq 2, has six GCOD descriptors, three of which, GC1(any), GC2(hbd), and GC3(p+), represent similar structural features as GC1′(hbd), GC2′(p+), and GC3′(p+), respectively, in the best RI model, eq 4. The other three GCOD terms, GC4(any), GC5(any), and GC6(p+), of the RD-4D-QSAR model involve occupancy by receptor residue atoms of Leu 136 and Ser 674. The RD-4D-QSAR model has an even distribution of descriptors arising from both the inhibitor and the receptor suggesting that both inhibitor and receptor should be included when conducting a QSAR analysis. Moreover, the explicit information about the receptor geometry embedded in the QSAR provides insight into conformational changes in the receptor due to a particular bound inhibitor.

Induced-fits as part of ligand–receptor binding has long been recognized. However, capturing ligand–receptor induced-fitting in structure-based design and/or QSAR analysis has proven elusive. Superimposing the conformations of Asn 284 for complexes 24, 5, and 10, as shown in Figure 12, clearly reveals that Asn 284 realigns itself to form optimum hydrogen bonding with each of the three bound inhibitors and also alters its role in each ligand binding process. In complex 24 Asn 284 acts as a hydrogen bond donor, whereas in complex 5, it is a hydrogen bond acceptor through residue repositioning and also in complex 10 but through a residue "flip-flop". Realignment of Asn 284 is essential for GPb-inhibitor recognition.

A composite view of the observed changes in the binding modes of the GPb binding site is graphically represented in Figure 13 using complexes 1 and 10, that is, two of the most potent training set inhibitors. Inhibitor 1 has only a β substituent, while inhibitor 10 has only an α substituent. Three major conformational changes are observed in Figure 13: (1) residue repositioning and flip-flopping of Asn 284, (2) the phosphate group of PLP moving into close vicinity to the β substituent of inhibitor 10, and (3) reorientation of the backbone carbonyl group of His 377 upon binding both inhibitors 1 and 10. When inhibitors of varing chemical structure bind to the receptor, conformational changes are induced in the key recognition residues so that the inhibitor is aligned to optimize inhibitor–receptor binding interactions. These binding mode "adjustments", as a function of inhibitor structure, are seemingly well-characterized by RD-4D-QSAR analysis. Moreover, these findings of induced-fit suggest that

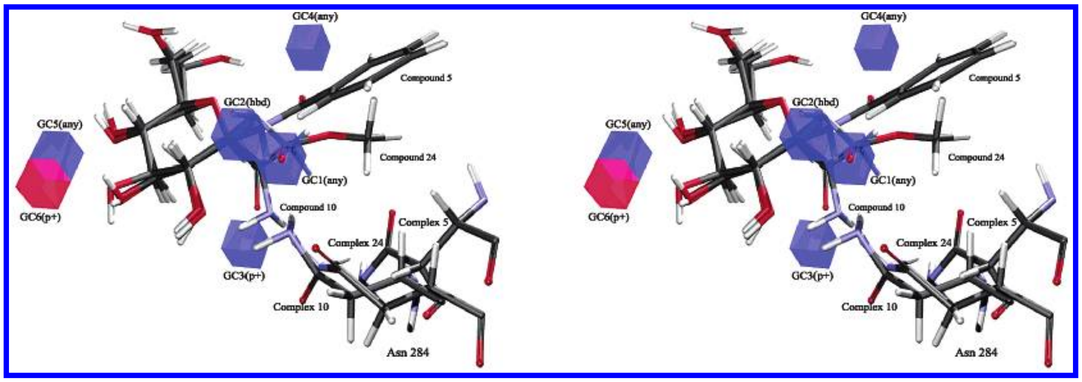


Figure 12. Conformation and alignment changes of Asn 284 as a function of inhibitor binding. Geometries of Asn 284 in complex 24, 10, 5, are superimposed.

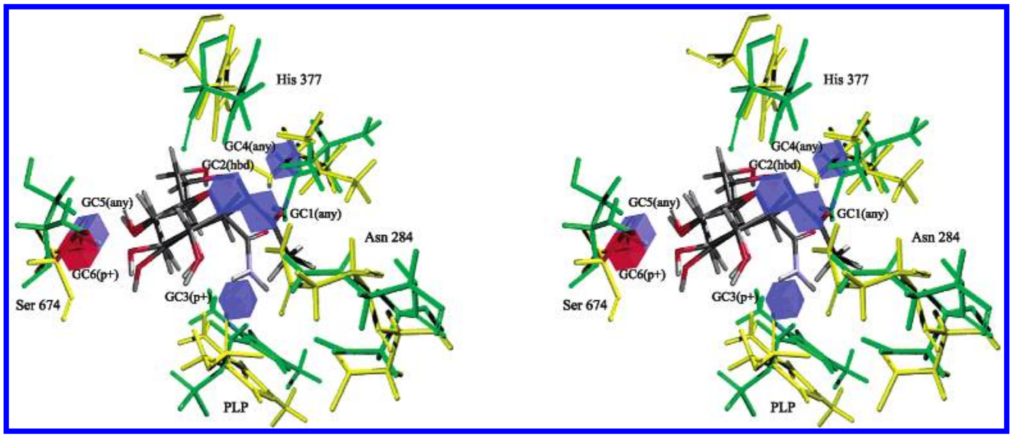


Figure 13. Induced ligand−receptor fits at the binding site of GPb. The conformation of complex 1 is green and complex 10 yellow. Three major differences are observed: (1) reorientation of the backbone −CO− of His 377, (2) residue flipping for Asn 284, and (3) realignment of the phosphate group of PLP.

Table 6. Test Set Predictions Using the Best RI-4D-QSAR Model and the Corresponding Descriptor Values for Each Test Set Compound

compd no	GC1′(0,3,2,hbd)	GC2′(1,−2,1,p+)	GC3′(−3,1,0,p+)	GC4′(−1,3,3,p+)	GC5′(2,−1,−1,np)	GC6′(−4,3,2,np)	pred	obs	res
T1	0	0	0	0	0.002	0	2.91	3.21	−0.30
T2	0	0	0	0	0.032	0.001	3.26	3.37	−0.11
T3	0	0	0	0	0	0	2.89	3.58	−0.69
T4	0.002	0.002	0	0.023	0.001	0	3.11	4	−0.89
T5	0	0.485	0	0	0	0	5.17	5.64	−0.47
T6	0.168	0.166	0	0	0	0.005	4.05	5.32	−1.27
T7	0.006	0.108	0	0	0	0	3.41	7.66	−4.25
T8	0	0.817	0.015	0	0	0	7.01	5.86	1.15

Table 7. Test Set Predictions Using the Best RD-4D-QSAR Model and the Corresponding Descriptor Values for Each Test Set Compound

compd no.	GC1(0,−3,2,any)	GC2(1,−2,2,hbd)	GC3(−3,−1,3,p+)	GC4(−1,−2,−3,any)	GC5(1,4,3,any)	GC6(1,4,4,p+)	pred	obs	res
T1	0.245	0.39	0	0	0.01	0.64	3.46	3.21	0.25
T2	0.13	0.265	0	0	0.125	0.71	3.08	3.37	−0.29
T3	0.005	0	0	0.065	0.035	0.04	3.64	3.58	0.06
T4	0.125	0	0.35	0	0.005	0.255	5.14	4	1.14
T5	0.11	0.28	0.315	0	0.075	0.69	4.97	5.64	−0.67
T6	0.02	0.115	0.74	0	0.025	0.715	6.87	5.32	1.55
T7	0.06	0.095	0.815	0	0.07	0.795	7.25	7.66	−0.41
T8	0.015	0.485	0.265	0	0.31	0.63	5.81	5.86	−0.06

caution should be taken in using a rigid X-ray, or even a semiflexible, protein structure in structure-based design studies as is common in current application.

The RI-4D-QSAR model, eq 4, has about the same statistical quality, with respect to the training set, as the RD-4D-QSAR model, eq 2. However, the RI-4D-QSAR model poorly predicts the more rigid compounds (T6, T7, T8) of the test set as compared to the RD-4D-QSAR model. Some insights into the source of the poor RI-4D-QSAR predictions can be gleaned from an inspection of Table 6. The high GC2′(p+) value, 0.82, for T8 and the low GC1′(hba) values, 0 and 0.006, for compounds T6 and T7, respectively, of eq 4 may be responsible for the poor predictions. These two GCODs identify the same intermolecular interactions as GC1(any) and GC2(hbd) of eq 2. However, GC2′(p+) and GC2(hbd) differ in specifying the relative locations of the binding interactions. In addition, GC1′(hbd) only captures the binding role of the proton of the inhibitor −NHCO− group. However, GC1(any) captures both the hydrogen donor and acceptor binding behavior of this inhibitor amide group.

The CEPs of the RI-4D-QSAR analysis are independent of geometrical information from the receptor. Thus, this CEP is necessarily a deviation from the actual more restricted conformational representation of the bound ligand and a proximate training set binding model. Furthermore, RI-4D-QASAR models cannot incorporate ligand−receptor induced-fit effects which may realign the binding site residues to create multiple binding modes even within a set of analogue ligands.

Predicted inhibition potency measures of the test set compounds and the corresponding GCODs occupancy values of the RD-4D-QSAR model (eq 2) are shown in Table 7 and can be directly compared to those of the RI-4D-QSAR model (eq 4). The occupancy values of the similar GCODs between the best RD- and RI- models are different from one another suggesting that the inhibitors experience different physiochemical environment in the simulation process in the RI and RD studies. However, despite the GCOD difference in RD- and RI- models, similar GCODs capture the same inhibitor−receptor interactions which demonstrates the strength and reliability of the 4D-QSAR analyses.

RD-4D-QSAR analysis diminishes the number of trial alignments that need to be considered as compared to RI-4D-QSAR. Basically, alignments which were reasonable in an earlier RI-4D-QSAR study[5] are sterically forbidden in the RD-4D-QSAR analysis because of overlaps of parts of a ligand with residues in the lining of the binding site. In some ways the degrees of RI-4D-QSAR alignment freedom, seemingly eliminated by doing RD-4D-QSAR analysis, really are not. The realignment of receptor residues realized in RD-4D-QSAR analysis, due to induced ligand−receptor fitting, can be thought of as a re-expression of the RI-4D-QSAR alignment freedom.

We have assumed that GCODs in a RI-4D-QSAR model with the "any" IPE type, and having negative regression coefficients (occupancy decreases activity), correspond to sites occupied by receptor atoms and forbidden to the ligand. This assumption has held up for those cases where a RI-4D-QSAR 3D-pharmacophore could be aligned within the binding site geometry of the target receptor. On the other hand, we have been stymied to interpret RI-4D-QSAR GCODs with IPE type "any" but with positive regression coefficients. An obvious interpretation might be that occupancy by "any" type of atom could (marginally) enhance binding through favorable Van der waal dispersion interations. However, in several instances the near-exclusive occupancy of these GCODs is a combination of "hba" and "hbd" IPE types, which would seem in conflict to one another. This RD-4D-QSAR study, and the observed realignment behavior of Asn 284, demonstrates how, in fact, such a GCOD populated by "hba" and "hbd" IPE types can occur. For some inhibitors Asn 284 serves as a hydrogen bond donor site and for others a hydrogen bond acceptor site. Thus, a RI-4D-QSAR model GCODs with IPE type "any", and positive regression coefficients, may be reflecting sites of receptor induced-fit.

Overall, this study demonstrates that RD-4D-QSAR analysis can perform as a quantitative structure-based design tool. It accomplishes this feat by efficiently sampling a completely flexible ligand−receptor system, exploring a vast set of possible binding pharmacophores, identifying the best, and packaging the results in a compact, quantitative, and easy to understand pharmacophore model.

ACKNOWLEDGMENT

Resources of the Laboratory of Molecular Modeling and Design at UIC and The Chem21 Group, Inc. were used in performing this work. An unrestricted financial gift from the Procter and Gamble Company is also gratefully acknowledged.

REFERENCES AND NOTES

(1) Santos-Filho, O. A.; Mishra, R. K.; Hopfinger, A. J. Free energy force field (FEFF) 3D-QSAR analysis of a set of Plasmodium falciparum dihydrofolate reductase inhibitors. *J. Comput. Aided Mater. Des.* **2001**, *15*, 787−810.

(2) Venkatarangan, P.; Hopfinger, A. J. Prediction of Ligand−Receptor binding thermodynamics by Free Energy Force Field three-dimensional quantitative structure−activity relationship analysis: applications to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *42*, 2169−2179.

(3) 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. Compu. Sci.* **1997**, *37*, 792−811.

(4) 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.

(5) 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*, 1141−1150.

(6) Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; Alencastro, 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.

(7) Ravi, M.; Hopfinger, A. J.; Hormann, R. E.; Dinan, L. 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1587−1604.

(8) Krasowski, M. D.; Hong, X.; Hopfinger, A. J.; Harrison, N. L. 4D-QSAR analysis of a set of propofol analogues: mapping binding sites for an anesthetic phenol on the $GABA_A$ receptor. *J. Med. Chem.* **2002**, *45*, 3210−3221.

(9) Santos-Filho, O. A.; Hopfinger, A. J. The 4D-QSAR paradigm: application to a novel set of nonpeptidic HIV protease inhibitors. *Quant. Struct.-Act. Relat.* **2002**, *21*, 369−381.

(10) Hong, X.; Hopfinger, A. J. 3D-pharmacophores of flavonoid binding at the benzodiazepine GABAA receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.

(11) Martin, J. L.; Veluraja, K.; Ross, K.; Johnson, L. N.; Fleet, G. W. J.; Ramsden, N. G.; Bruce, I.; Orchard, M. G.; Oikonomakos, N. G.; Papageorgiou, A. C.; Leonidas, D. D.; Tsitoura, H. S.; Glucose analogue inhibitors of glycogen phorphorylase: the design of potential drugs of diabetes. *Biochemistry* **1991**, *30*, 10101−10116.

(12) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Bichard, C. J. F.; Orchard, M. G.; Fleet, G. W. J.; Oikonomakos, N. G.; Leonidas, D. D.; Kontou, M.; Papageorgiou, A. C. Design of inhibitors of glycogen phosphorylase: a study of α- and *β*-C-glucosides and 1-thio-*β*-D-glucose compounds. *Biochemistry* **1994**, *33*, 5745−5758.

(13) Watson, K. A.; Mitchell, E. P.; Johnson, L. N. Glucose analogue inhibitors of glycogen phosphorylase: from crystallographic analysis to drug prediction using *GRID* force field and *GOLPE* variable selection. *Acta Crystallogr*. **1995**, *D51*, 458−472.

(14) Pastor, M.; Cruciani, G.; Clementi, S. Smart region definition: a new way to improve the predictive ability and interpretability of three-dimensional quantitative structure−activity relationship. *J. Med. Chem.* **1997**, *40*, 1455−1464.

(15) Martin, J. L.; Johnson, L. N.; Withers, S. G. Comparison of the binding of glucose and glucose 1-phosphate derivatives to T-state glycogen phosphorylase b. *Biochemistry* **1990**, *29*, 10745−10757.

(16) http://www.rcsb.org/pdb/.

(17) 4D-QSAR Program, Version 3.0, The ChemBats21 Group Inc., Lake Forest, IL, 2001.

(18) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T. An all atom force field for simulation of proteins and nucleic acids. *J. Comput. Chem*. **1986**, *7*, 230−252.

(19) HyperChem Program Release 5.01 for Windows; Hypercube, Inc., 1996.

(20) HyperChem Reference Manual, Hypercube, Inc., 1996.

(21) Doherty, D. C. *MOLSIM User's Guide*; The ChemBats21 Group, Inc.: Lake Forest, IL, 1997.

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

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

(24) Friedman, J. *Multivariate adaptive regression splines*; Technical Report No. 102; Laboratory for computational statistics, Department of Statistics, Stanford University: Stanford, CA, 1988; reversed 1990.

(25) Kleinbaum, D. G.; Kupper, L. L.; Muller, K. E.; Nizam, A. *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed.; Books/Cole Publishing Company: Pacific Grove, CA, 1998; pp 330−332.

(26) SAS, Version 8.1 for Windows, SAS Institute Inc., 2001.

(27) Karpus, M.; Mccammon, J. A. *Annu. Rev. Biochem*. **1983**, *53*, 263.

CI0340714