



Evaluation of designed ligands by a multiple screening method: Application to glycogen phosphorylase inhibitors constructed with a variety of approaches

Sung-Sau So^{a,*} & Martin Karplus^{a,b,**}

^a*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, U.S.A.*; ^b*Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France*

Received 24 November 2000; accepted 11 May 2001

Key words: binding affinity prediction, glycogen phosphorylase inhibitor, LUDI, MCSS, QSAR, structure-based drug design

Summary

Glycogen phosphorylase (GP) is an important enzyme that regulates blood glucose level and a key therapeutic target for the treatment of type II diabetes. In this study, a number of potential GP inhibitors are designed with a variety of computational approaches. They include the applications of MCSS, LUDI and CoMFA to identify additional fragments that can be attached to existing lead molecules; the use of 2D and 3D similarity-based QSAR models (HQSAR and SMGNN) and of the LUDI program to identify novel molecules that may bind to the glucose binding site. The designed ligands are evaluated by a multiple screening method, which is a combination of commercial and in-house ligand-receptor binding affinity prediction programs used in a previous study (So and Karplus, *J. Comp.-Aid. Mol. Des.*, 13 (1999), 243–258). Each method is used at an appropriate point in the screening, as determined by both the accuracy of the calculations and the computational cost. A comparison of the strengths and weaknesses of the ligand design approaches is made.

Introduction

Glycogen is a readily mobilized reserve of glucose, the major fuel for maintaining cellular metabolism in most mammals. Glycogen phosphorylase (GP) catalyzes the phosphorolysis of glycogen main-chain (α -1,4-glucoside)_n to glucose-1-phosphate (Glc-1-P), an intermediate metabolic product in glycogen degradation. In muscle, Glc-1-P is consumed via glycolysis to sustain muscular contraction, and in the liver, it is converted to α -D-glucose and released into the bloodstream as a source of energy for other tissues [1]. GP also has an important regulatory role in the glycogen metabolism owing to its weak binding affinity with glucose. At a high concentration, glucose acts as a

competitive inhibitor with the product Glc-1-P and induces a conformational change in the enzyme structure from the *R* state to the *T* state. The allosteric mechanism facilitates the catalytic dephosphorylation of GP by the protein phosphatase 1 (PP1), which results in the conversion of GP from its active *a* form to its inactive *b* form. This shift in equilibrium also leads to the activation of other enzymes, notably glycogen synthase that promotes the synthesis of glycogen.

There has been considerable interest in designing a more potent GP inhibitor than its physiological regulator, α -D-glucose. Such compounds could be useful as therapeutic agents in the treatment of diabetes, a disease of great importance in many developed countries. Previous studies of GP inhibitors have been focused on glucose analogs, where the structural variations are limited to the α - and β -substituents of the anomeric C1 atom [1, 2]. The best known GP inhibitor (GP3

*Present address: Hoffmann-La Roche Inc., 340 Kingsland Street, Nutley, NJ 07110, U.S.A.

**To whom correspondence should be addressed: E-mail: marci@tammy.harvard.edu

in Figure 1) to date has a K_i value of 3 μM , which is more potent than the parent glucose molecule (GP1 in Figure 1; $K_i = 1.7 \text{ mM}$) by nearly three orders of magnitude. However, to mimic the regulatory effect of glucose at its physiological concentration, it has been estimated that a potential drug candidate needs to reach a potency (K_i) in the range of 0.10–0.01 μM [1]. Such an improvement by two orders of magnitude over the best available compound is a challenging problem in ligand design.

Computer-aided ligand design strategies can be categorized into three principal types [3, 4]. The first is the so-called data-mining method, which involves the virtual screening of a library of chemicals, typically corporate databases or collections of commercially available compounds. The search is usually based on a pharmacophore model or a similarity-based key that has been established by a few lead ligands [5]. A pharmacophore model defines spatial inter-relationships between the sites characterizing the key features of ligand-protein binding interactions (e.g. hydrogen bond donors and acceptors, positive and negative charge centers, aromatic rings and hydrophobic groups). During the database search, the conformation space (either pre-compiled with low energy conformations or generated on-the-fly) of a given compound will be explored and the search algorithm determines whether the spatial constraints of the pharmacophore can be satisfied [6, 7]. A similarity-based search key makes use of molecular similarity to screen new leads from databases. The similarity descriptions may be derived from atom-based features (e.g., atomic properties, important functional groups and substructures, or molecular connectivity indices) or 3D field-based descriptions of the steric, electrostatic or hydrophobic microenvironment of the ligands. The second ligand design approach employs an active analog approach to optimize the structures of existing lead compounds [8, 9]. Empirical QSAR models are usually the theoretical basis of structural optimization since the target structure is not known in many cases. The third approach exploits structural information about the protein, which provides a three-dimensional environment to dock molecular fragments where they can be subsequently connected to form *de novo* ligands; or to grow molecules that fit the binding site in a complementary manner. Many drug design programs have been developed to perform such tasks [10–25].

In this paper we employ a number of ligand design techniques to design better regulators of GP that

can shift the balance from glycogen degradation to glycogen synthesis. One purpose of this study is to compare a variety of different approaches that have been proposed for ligand design. We attempt structure optimization of existing GP inhibitors with new substituents attached to the glucose moiety. This is done by manual model building based on functional group maps obtained by multiple copy simultaneous search (MCSS) [16, 17, 26] or molecular fragments identified by LUDI [12], or an automated procedure based on a CoMFA [27] QSAR model. Also, we try to discover new leads that are derived from entirely different chemical classes. Short peptide sequences are constructed using a combination of the MCSS and OLIGO [28] programs. A number of non-glucose analogs are identified by database searches using 2D fragment-based or 3D similarity-based queries. In three studies, the actual protein binding site structure provides the basis for the placement of commercially available chemicals, a pre-selected database of small molecules or combinatorial library products. As a control, 30 compounds are taken at random from the National Cancer Institute (NCI) library to provide a reference point for evaluating the effectiveness of the various design approaches.

An accurate means to estimate the affinity of putative protein-ligand complexes is critical to the success of any computer-aided molecular design strategy. In the past years, a large number of empirical scoring functions have been formulated for this purpose. Some of them are high-throughput (e.g., LUDI) and are applicable for the rapid screening of large databases or virtual combinatorial chemistry libraries. A recent example is the scoring function proposed by Muegge and Martin based on potentials of mean force [29]. Others require significant computational resources and are therefore restricted to the detailed examinations of a few molecules [30]. Another interesting development in this area is the notion of consensus scoring, which was proposed by Murcko and co-workers [31]. They performed an extensive docking study in which thirteen different scoring functions were used to evaluate ligand-protein interactions for three enzyme targets. The combined use of many scoring functions led to a substantial reduction in the number of false positives. In the present study, we introduce a multi-layer approach for predicting the activity of the designed compounds. In a previous paper [32], we examined 30 known GP ligands with seven prediction methods. The results of the study suggested that a combination of these prediction methods could lead to an effective

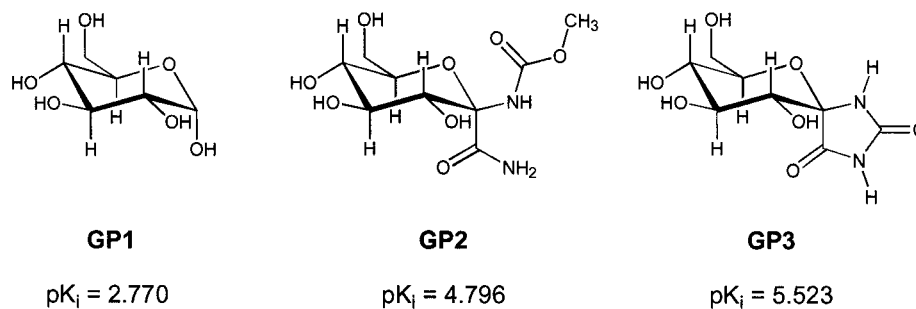


Figure 1. The three GP inhibitors that have been used as design template in this study. The experimental pK_i values of **GP1-3** are also shown.

scoring function. In this paper, we use this system-specific scoring system to estimate the binding affinity of proposed new ligands for GP.

Methods

I. Activity prediction

The seven activity prediction methods used in the previous paper [32] included five QSAR-based models (HQSAR [33], CoMFA [27], SMGNN [34, 35], C2GNN [36, 37] and RSM [38, 39]) and two structure-based methods (LUDI [12] and SBEP [32]). Since different methods involve different sets of assumptions and limitations, some methods should provide more accurate predictions for certain types of molecules than others. This suggests that instead of employing only the method that yielded the highest correlation in the GP ligand evaluation, it would be better to create a consensus scoring function by combining the various prediction methods in a logical way.

We describe here a multi-layer model for the activity prediction of GP ligands. The grouping of the various prediction methods is made on the basis of the individual results and their associated computational cost. The details of each of the layers are described below.

(a) *Screening layer*

The first layer is a preliminary screening layer that can evaluate a large number of compounds at minimal computational cost. For this part of the calculation, accurate numerical predictions of activity are not essential. What one needs to do is to exclude compounds that are likely to be poor binders. The LUDI program, which is commonly used for screening large databases, appears to afford a good balance between

computational efficiency and prediction accuracy. Although in the previous study we found no significant correlation between LUDI scores and experimental pK_i for the 30 known GP inhibitors [32], their individual scores (all greater than 379) provide a baseline that a reasonable GP inhibitor would expect to attain. We decided to set a score of 300 to be the minimum threshold for a putative GP ligand. The time of the calculation is of the order of a CPU second per compound on a 175 MHz R10000 Silicon Graphics workstation.

(b) *Prediction layer*

The second layer is an activity prediction layer, whose principal function is to provide a more accurate assessment of biological activity. In this layer, we rely on QSAR models that are specifically parameterized for GP and should therefore give more reliable predictions for GP ligands than the LUDI program, which is calibrated by use of many classes of protein-ligand complexes. In the previous study, we described five QSAR models (HQSAR, CoMFA, SMGNN, C2GNN and RSM) which have good predictivity and yield q^2 values ranging from 0.60 to 0.82. Furthermore, a jury method that combines their individual predictions led to an increase in predictivity to a q^2 value of about 0.86 [32]. Thus, in this prediction layer, the activity prediction of each compound is given by the average value of the five QSAR assessments. The calculations in the layer take less than 1 CPU minute per compound.

(c) *Validation layer*

The final layer of the scoring system is provided by a structure-based binding energy predictor (SBEP), which serves as a validation of the activity prediction. The main reason for leaving this method to the final stage is that the calculations involved are substantially more expensive (approximately 2 CPU hours per compound) than the other methods. This model is

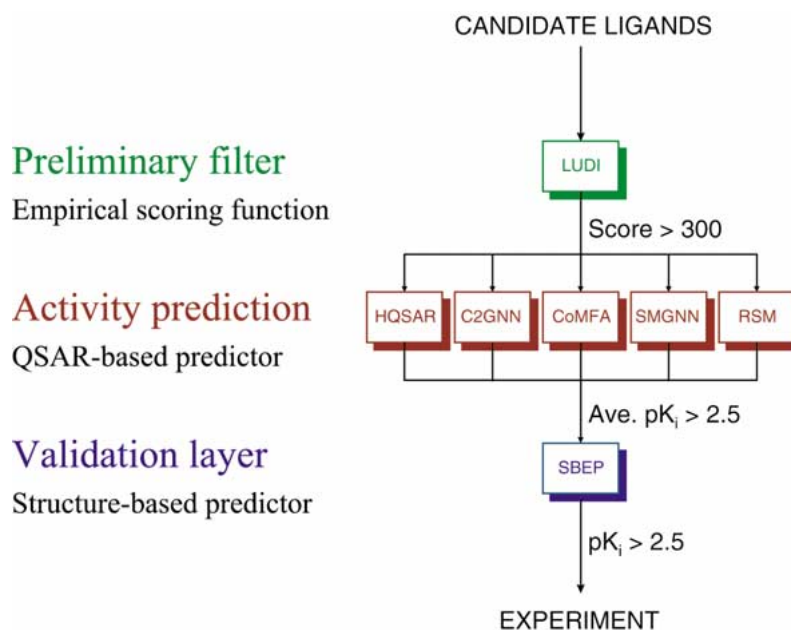


Figure 2. The multi-layer hybrid prediction system proposed in this study. The preliminary screening layer consists of the LUDI scoring function that can screen a large number of compounds very quickly. The activity prediction layer makes use of five QSAR-based models to provide a more accurate measure of activity. The final validation layer takes advantage of the 3D ligand-protein structures and generates thermodynamics properties that are derived from atomistic potential energy functions for prediction. A candidate ligand must pass the specified criteria for each of the three layers before further experimental work is carried out.

based on parameters that describe protein-ligand binding in atomistic details. They are Int_{vdw} and Int_{elec} , the van der Waals and electrostatic interaction energies as calculated in vacuum; ΔG_{elec} , the electrostatic contribution to the free energy of binding; and $T\Delta S_{\text{rec}}$, the entropy loss of the protein due to ligand binding. In a preliminary analysis, a four-descriptor linear regression model, which yielded a q^2 value of 0.51, was found. It is:

$$\text{pK}_i = 5.557 - 0.103\text{Int}_{\text{elec}} - 0.238\text{Int}_{\text{vdw}} - 0.076\Delta G_{\text{elec}} + 0.246T\Delta S_{\text{rec}}. \quad (1)$$

The equation predicts that stronger intermolecular interactions (Int_{elec} and Int_{vdw}) between ligand and receptor will increase inhibitory activity. Since the two interaction energy terms are always negative, they contribute positively to the binding. On the other hand, the ΔG_{elec} (always positive) and the $T\Delta S_{\text{rec}}$ (always negative) terms, as expected, weaken the binding of the ligand. Overall, the signs for the four coefficients are consistent with the physical attributes of ligand binding. Later, a neural network was used to obtain optimal nonlinear combination of these properties, leading to a significant increase in predictivity of the SBEP model

($q^2 = 0.61$). The functional forms of the descriptors, as depicted by a neural network monitoring scheme [40, 41], were consistent with the regression analysis results. For details of the calculation, readers are referred to the previous publication [32].

The overall scoring scheme is summarized in Figure 2. First the LUDI program was used as a preliminary screen to remove the less promising candidates, where a minimum LUDI score of 300 was required for a candidate ligand to proceed to the next prediction phase. The prediction layer, which reports the average predicted pK_i value from five different QSAR-based approaches, provided a more reliable activity assessment at low computational cost. Ligand candidates with predicted pK_i values higher than 2.5 would enter the final validation layer, where prediction based on atomic force field and electrostatic calculations between the protein-ligand complex were calculated. The same activity criterion (SBEP $\text{pK}_i > 2.5$) was used to determine the final selection of the ligands. Finally, the ligands were ranked according to

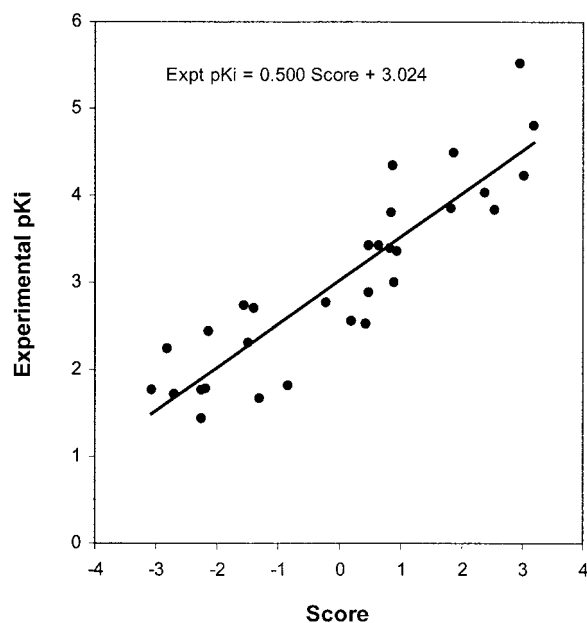


Figure 3. The experimental pK_i values versus the scores of the 30 training compounds. The scores are derived from Equation 2 based on the calculated pK_i values.

a composite score given by the following expression:

$$\text{Score} = \frac{\text{QSAR}_{\text{pred}} - \text{Average}(\text{QSAR}_{\text{pred, training set}})}{\text{Standard Deviation}(\text{QSAR}_{\text{pred, training set}})} + \frac{\text{SBEP}_{\text{pred}} - \text{Average}(\text{SBEP}_{\text{pred, training set}})}{\text{Standard Deviation}(\text{SBEP}_{\text{pred, training set}})} \quad (2)$$

This score is derived from the predicted pK_i values of the combined QSAR model in the prediction layer and the SBEP model in the validation layer. Each score component is standardized with respect to the range of predicted pK_i activities for the 30 training compounds used in the QSAR (1.73 to 4.66) or SBEP (1.55 to 4.55) models [32]. The use of standardized values reduces the bias that may be caused by unusually large variation of predicted values from a given method. For the present case, the average and standard deviation, as well as the range of the QSAR (3.04 and 0.88) and SBEP (3.03 and 0.86) pK_i predictions for the 30 training compounds were very similar. Based on equation 2, the scores for the training compounds are calculated and they range from -3.08 to 3.18 . The correlation coefficient between these scores and the experimental pK_i values is very high ($r = 0.90$; see Figure 3), and the two quantities can be related by the following expression:

$$pK_i(\text{experimental}) = 0.500 \times \text{Score} + 3.024 \quad (3)$$

The composite score given by equation 2 can be converted to a final predicted pK_i ($pK_{i,\text{final}}$) for each compound using the relationship above.

II. Ligand design

This section describes the ligand design methods used in this study. Each method is identified by a double header. The first header specifies the method that was used for the positioning of fragments (or whole molecules) in the binding site or the selection of new molecules. This includes structure-based MCSS, LUDI methods, and ligand-based CoMFA, HQSAR, FieldFit and SMGNN approaches. The second header (after the slash) identifies the source of the new ligand molecules. They could come from an easily accessible source such as a combinatorial chemistry library (CombiChem), peptide system (OLIGO) or existing molecular databases. Three databases were used in this study, the available chemical directories (ACD), the National Cancer Institute (NCI) library, and a relatively small database (SDB) of organic molecules. We tested certain databases with particular methods simply as a matter of convenience, i.e., the standard database that had a built-in interface with the method under consideration was used. Other ligand candidates were novel compounds that were made by linking suitable fragments to pre-defined template molecules (LINK). A relatively simplistic linking strategy was used in this study. To restrict the search space, we only considered connections where the whole or part of the fragments could be attached directly to the templates without adding new linker atoms [16].

(a) MCSS/LINK

The multiple copy simultaneous search (MCSS) program [14–16, 26] was used to create functional group maps of the binding site of glycogen phosphorylase (GP). In a run that was typical in this study, 500 replicas of a given function group were randomly distributed over a sphere of 12 \AA radius centered at the binding site (the coordinates of the C1 atom of the glucose molecule). The replicas were then simultaneously, and independently, energy minimized in the field of the rigid GP structure, such that they interacted with the protein but not with each other. GATHER, an auxiliary utility program of MCSS, was invoked after minimization to remove duplicate minima based on a rms cutoff of 0.2 \AA . The above protocol was repeated ten times for each of the functional groups. The functional groups that were investigated in this study are listed in

Figure 4. The polar hydrogen representation (parameter set 19 [42] in the CHARMM program) was used for both the protein and the functional groups.

As an initial test we compared the methanol (MEOH) functional map to the binding conformation of α -D-glucose. Figure 5 shows all the MEOH MCSS minima that are within 1.5 Å to the glucose molecule. Three of the MCSS minima overlay the hydroxyl groups (O3, O4 and O6) in glucose, reproducing almost exactly both the location and orientation of the OH bonds. No MEOH minima were found near the O1 and O2 hydroxyl groups of the glucose molecule. This suggests that hydroxyl groups are not the optimal choice at these two positions. This is well-known for O1, the focus of structural modification of the existing series of GP inhibitors, where the replacement of this hydroxyl group by alternative chemical groups have led to analogs with significantly improved potency [1, 2]. It is a more surprising result for O2 since crystallographic study of the complex has indicated that the latter group forms adequate hydrogen bonds with the protein [1, 2]. However, no studies to modify this hydroxyl group or free energy simulations to determine the contribution of this group (e.g., simulations that changes OH to H) to the binding site have been performed. We have investigated the possibility of replacing the C2 hydroxyl group with other substituent groups found by MCSS.

We constructed candidate ligands by linking some of the MCSS minima to three template molecules: the parent glucose molecule (Figure 1: **GP1**) and the two most active analogs (Figure 1: **GP2** and **GP3**) so far available. To simplify the task, we made direct attachments (i.e., no linker groups) so that only the MCSS minima that were within 2 Å of the templates were examined. The connection of the MCSS groups to the template scaffold were made based on a visual analysis of the orientation of the functional groups relative to the template. The new bond was constructed so as not to introduce significant internal strain in the candidate ligand (i.e., the bond length, bond angles and dihedral angles were required to be near their minimum energy values). The newly formed ligand molecules were subsequently energy minimized in the rigid protein to regularize the internal coordinates. Six compounds (Table 1(a): **A1-6**) were suggested by this analysis. It should be noted that it is not necessary to attach the entire MCSS group to the template. For example, **A2** was constructed by linking the C₈ atom of a LYSR group to the C1 atom of **GP1**; thereafter a few atoms of the original LYSR group were removed to re-

lieve unfavorable intramolecular contacts (C_β and C_γ) and to eliminate positive charge (3H_c). In a number of cases, the hydroxyl groups in the templates were changed to secondary amine group to accommodate the substituents (e.g., **A5-6**).

Earlier crystallographic studies have identified a bound water molecule (wat897) in the crystal structure [1]. This molecule is part of a hydrogen bond network between GP and the O4 atom of the glucose analogs. We constructed an extended substituent group that replaces this bound water, while preserving the overall hydrogen bonding pattern. Three compounds (**A7-9**) are suggested with the bound water incorporated into the ligand molecule. The bound water was replaced by a hydroxyl group that was attached to a five-member ring connecting the C3 and C4 atoms of the glucose template.

(b) MCSS/OLIGO

The OLIGO [28] program automatically creates oligomeric ligands using the positions of constituent fragments. It was applied to construct peptidic ligands based on the MCSS functional maps obtained previously (in (a)). OLIGO joined a number of *N*-methylacetamide (NMAC) MCSS minima to form a peptide backbone, and functional groups corresponding to different amino acid (AA) side-chains were attached to the backbone if a reasonable C_α-C_β bond length could be maintained. Specifically the calculation is based on a simulated annealing algorithm using a simplified energy function that penalized deviations from idealized bond length and intramolecular contact in the candidate oligomers. Thus, the procedure of OLIGO is similar to the CONNECT program [15], though they differ in two major ways: (a) for main chains generation, OLIGO uses simulated annealing whereas CONNECT uses a branch-and-bound algorithm [43]; (b) the penalty function for steric contacts is a repulsive electrostatic term in OLIGO whereas a quadratic function that depends only on distance is implemented in CONNECT. After the bond connections were made, the conformation of the oligomer was refined by energy minimization in the fixed protein, and the specific stereochemistry of the C_α atoms was introduced by applying an appropriate improper energy term corresponding to an L- or a D-peptide. In this method we considered peptides with all L or all D configurations; for future work an improved implementation that generates oligomers with mixed configuration by assigning a particular stereo-center based on the sign of the C_{α,main}-N_{main}-C_{main}-C_{β,side} im-

Table 1(a). Nine ligands that are suggested by the MCSS/LINK approach. The parent structures where the compounds are derived from are given in parenthesis. The overall score of each compound is computed from the QSAR and SBEP activity predictions (Equation 2), which is then converted to a predicted pK_i ($pK_{i,final}$) value using the relationship shown in Equation 3. The $pK_{i,final}$ of a compound is given in bold typeface if it passes the preliminary filter criteria (i.e., LUDI score > 300).

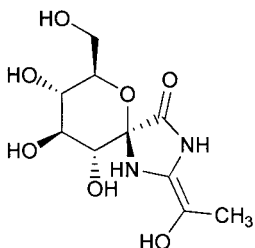
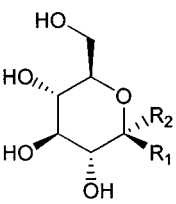
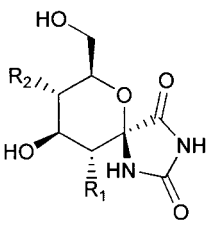
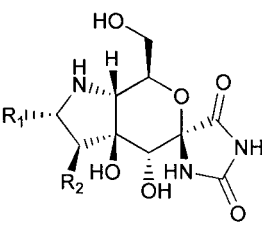
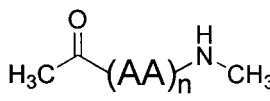
<div style="display: flex; justify-content: space-around; align-items: flex-end;"> <div style="text-align: center;">  <p>A (GP3)</p> </div> <div style="text-align: center;">  <p>B (GP1)</p> </div> <div style="text-align: center;">  <p>C (GP3)</p> </div> <div style="text-align: center;">  <p>D (GP3)</p> </div> </div>				
no.	structure	R ₁	R ₂	$pK_{i,final}$
A1	A			3.77
A2	B	H	CH ₂ CH ₂ NH ₂	2.34
A3	C	NHNHCH ₃	OH	3.06
A4	B	CH ₂ SCF ₃	H	1.90
A5	C	NHCH ₂ CF ₃	OH	1.85
A6	C	OH	NHCH ₂ OH	3.22
A7	D	OH	H	3.89
A8	D	OH	OH	4.19
A9	D	OH	NH ₂	3.82

Table 1(b). Six tri-peptides and eight tetra-peptides obtained using the MCSS/OLIGO method.

						
no.	n	AA1	AA2	AA3	AA4	$pK_{i,final}$
B1	3	L-Gln	L-Thr	L-Gln		1.35
B2	3	L-Ser	L-Thr	L-Gln		1.35
B3	3	D-Thr	D-Asn	D-Ser		1.26
B4	3	D-Ser	D-Gln	D-Gln		1.42
B5	3	D-Thr	D-Thr	D-Gln		1.42
B6	3	D-Thr	D-Thr	D-Ser		1.29
B7	4	L-Thr	L-Thr	L-Ser	L-Ser	1.24
B8	4	L-Thr	L-Thr	L-Gln	L-Gln	1.12
B9	4	L-Thr	L-Thr	L-Gln	L-Ser	1.16
B10	4	D-Thr	D-Thr	D-Ser	D-Ser	1.35
B11	4	D-Gln	D-Thr	D-Gln	D-Ser	1.55
B12	4	D-Ser	D-Gln	D-Thr	D-Gln	1.40
B13	4	D-Thr	D-Gln	D-Thr	D-Gln	1.28
B14	4	D-Ser	D-Gln	D-Gln	D-Thr	1.43

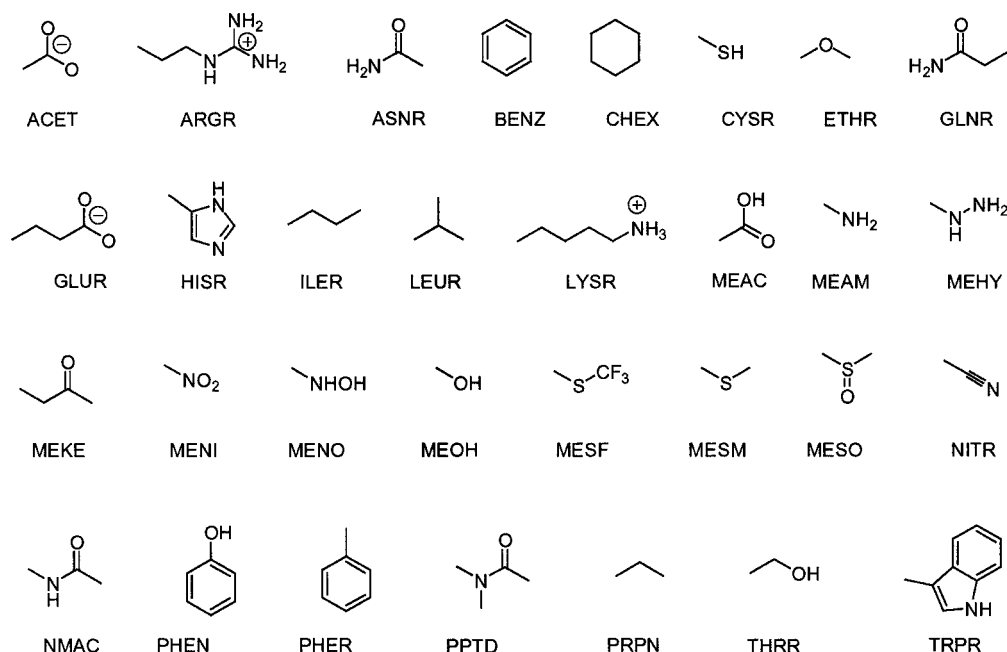


Figure 4. MCSS groups that are considered for this study.

proper dihedral angle could be used. In addition, the oligomers were limited to the four polar amino acid monomer types (Asn, Gln, Ser and Thr) due to the large number of potential hydrogen bond donor and acceptor groups that are present in the binding site. Standard blocking groups were patched to both the N- (Ace; CH_3CO) and the C-termini (Cbx; NHCH_3) [15]. Results from preliminary calculations suggest that oligomers up to 4 amino acids in size could be readily accommodated in the binding pocket. The MCSS groups that were used as input to OLIGO are summarized in Table 2.

In this method a number of tri- and tetra-peptide candidates were constructed. From 100 simulated annealing trials, 33 tri-peptide and 46 tetra-peptide sequences were generated. Two oligomers with opposite configurations (D versus L) were produced from each sequence, and their structures were regularized by energy minimization. The procedure led to a total of $2 \times (33 + 46) = 158$ candidate oligomers. Fourteen oligomers (Table 1(b): **B1–B14**) were selected from this pool based on the following criteria: (i) the planarity of the peptide bond must be maintained (deviation of the ω angle less than 30°); (ii) good hydrogen bonds with the binding site must be present. Figure 6 shows the construction of a tri-peptide (**B4**) from a number of MCSS functional group minima. The most com-

mon AA monomers were Thr, occurring in 20 cases, while the less bulky Ser was chosen on 12 occasions. Gln monomers were selected 17 times whereas the homologous Asn unit had only one occurrence (**B3**). This may be due to the repulsive electrostatic term in OLIGO, which typically gives a smaller penalty for Gln relative to Asn because its polar amide group is further away from the main-chain atoms. Thus, a connection with Gln would be preferred over Asn in situations where the difference between the $\text{C}_\alpha\text{--C}_\beta$ bonding energies was negligible. An interesting note is that both the L- and the D-form (**B7** and **B10**) of the sequence Ace-Thr-Thr-Ser-Ser-Cbx appeared to fit the binding site of GP.

(c) LUDI/ACD

LUDI is a *de novo* ligand design program for proteins [12, 13, 44, 45]. It generates interaction sites based on an empirical set of rules that are intended to determine favorable orientations for hydrogen bond (HB) formation and lipophilic contacts. Four different types of interaction sites are defined in the program: lipophilic-aliphatic, lipophilic-aromatic, hydrogen donor, and hydrogen bond acceptor. To perform a 3D search, a candidate molecule is superimposed on these interaction sites and is evaluated using a scoring function that is supposed to be related to the binding constant to protein.

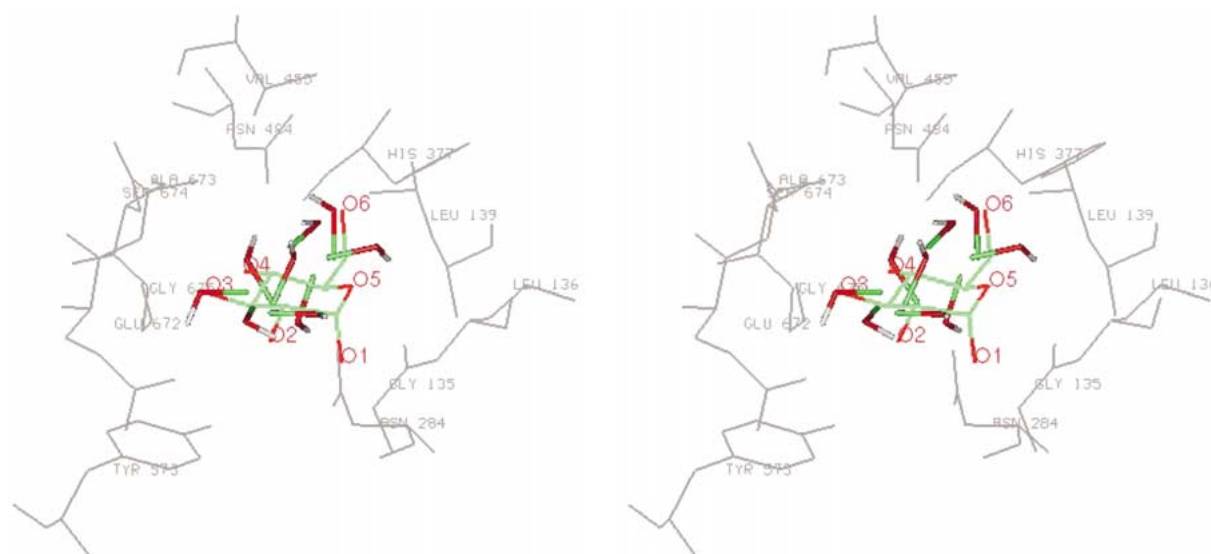


Figure 5. Stereopair diagram comparison of the methanol (MEOH) MCSS minima with the experimental binding conformation of glucose. All MEOH MCSS minima that are within 1.5 Å of the glucose molecule are shown. Three MEOH minima overlay well with the O3, O4 and O6 hydroxyl groups of glucose, but no corresponding minima are near O1 or O2.

Table 1(c). Ten hits derived from the LUDI/ACD approach. The calculated LUDI scores are also listed.

<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;"> A </div> <div style="text-align: center;"> B </div> <div style="text-align: center;"> C </div> <div style="text-align: center;"> D </div> <div style="text-align: center;"> E </div> <div style="text-align: center;"> F </div> </div>							
no.	structure	R ₁	R ₂	R ₃	R ₄	LUDI score	pK _{i,final}
C1	A	CH ₂ CO ₂ ⁻				508	2.93
C2	A	H				440	3.05
C3	B	NHCONH ₂	OH	H	H	430	3.64
C4	B	CH ₃	OH	OH	H	422	3.41
C5	C	NH				418	2.85
C6	D					411	2.63
C7	E					406	2.87
C8	F					399	2.33
C9	B	NH ₂	CO ₂ ⁻			392	2.02
C10	C	CH ₂				390	2.97

Table 1(d). Forty-four ligands constructed using the LUDI/LINK method.

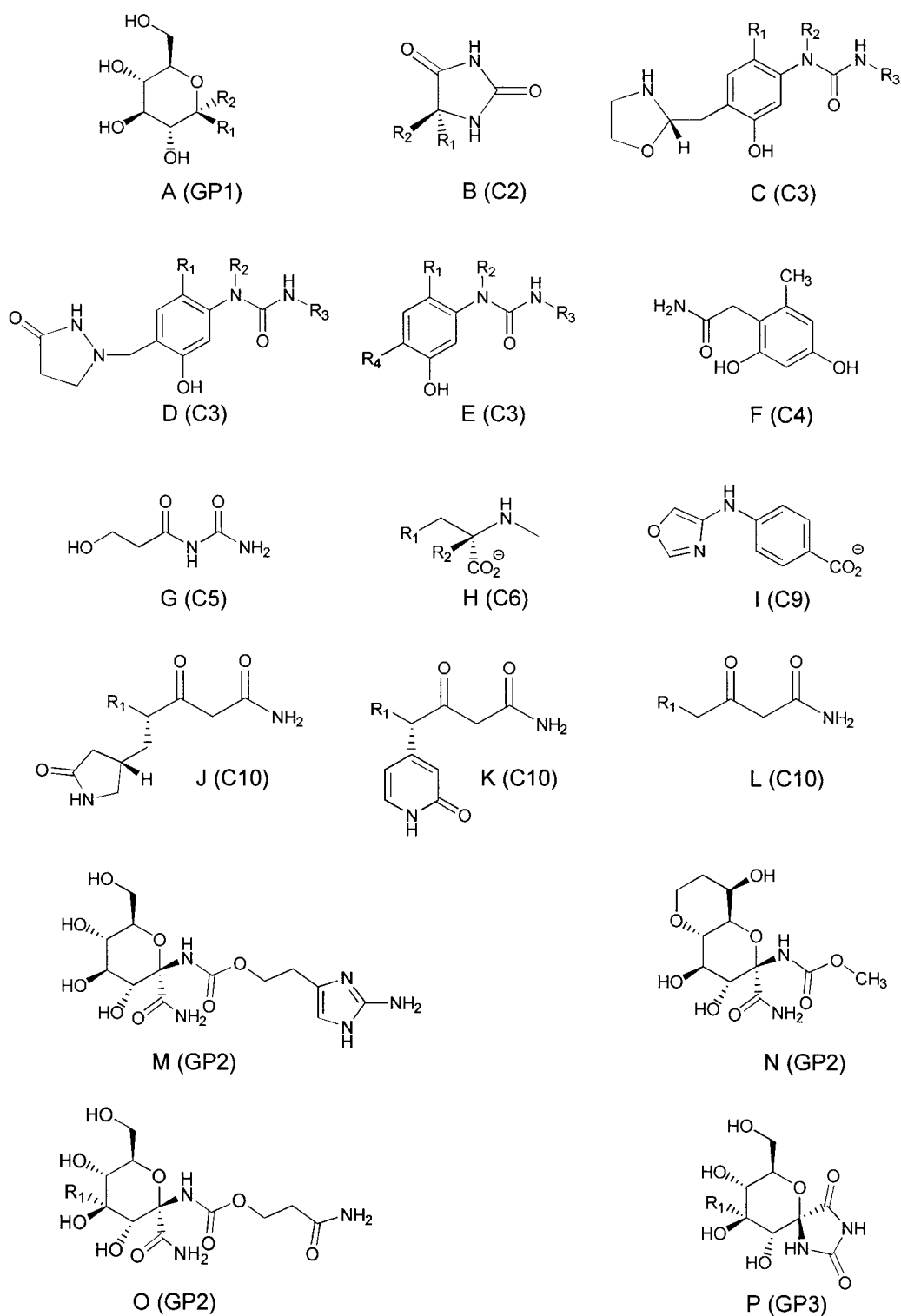


Table 1(d). Continued.

no.	str.	R ₁	R ₂	R ₃	R ₄	pK _{i,final}
D1	A	CH ₂ CH ₂ CH ₂ OH	H			3.04
D2	A	CH ₂ CH ₂ CH ₂ OH	CH ₂ OH			2.57
D3	A	H	CH ₂ NHCH ₃			2.47
D4	A	H	CH ₂ OH			3.44
D5	B	CH ₂ OH	CH ₂ CO ₂ ⁻			2.72
D6	B	CH ₂ CH ₂ CH ₂ OH	H			2.66
D7	B	CH ₂ CH ₂ OH	H			2.76
D8	C	H	H	H		2.78
D9	C	CH ₂ CH ₂ NH ₂	H	H		1.76
D10	C	CH ₂ CH ₂ NH ₂	CH ₂ OH	H		1.57
D11	C	CH ₂ CH ₂ NH ₂	CH ₂ OH	CH ₂ CHO		1.55
D12	C	CH ₂ CH ₂ NH ₂	H	CH ₂ CHO		1.58
D13	C	H	CH ₂ OH	CH ₂ CHO		1.79
D14	D	H	H	H		2.04
D15	D	CH ₂ CH ₂ NH ₂	CH ₂ OH	H		1.62
D16	D	CH ₂ CH ₂ NH ₂	CH ₂ OH	CH ₂ CHO		1.42
D17	D	CH ₂ CH ₂ NH ₂	H	CH ₂ CHO		1.48
D18	D	H	CH ₂ OH	H		1.83
D19	D	H	CH ₂ OH	CH ₂ CHO		1.65
D20	E	CH ₂ CH ₂ NH ₂	CH ₂ OH	CH ₂ CHO	H	1.59
D21	E	H	H	H	CH ₂ CH ₂ CONH ₂	3.39
D22	E	CH ₂ CH ₂ NH ₂	CH ₂ OH	CH ₂ CHO	CH ₂ CH ₂ CONH ₂	1.70
D23	E	CH ₂ CH ₂ NH ₂	H	CH ₂ CHO	CH ₂ CH ₂ CONH ₂	1.44
D24	E	H	CH ₂ OH	CH ₂ CHO	CH ₂ CH ₂ CONH ₂	2.31
D25	F					1.66
D26	G					2.84
D27	H	CH ₂ NHCH ₃	H			1.56
D28	H	CH ₂ NHCH ₃	CH ₂ CH ₂ OH			1.45
D29	H	H	CH ₂ CH ₂ OH			2.46
D30	I					3.33
D31	J	H				1.74
D32	J	CH ₂ NHCH ₃				1.63
D33	J	CH ₂ OH				1.74
D34	K	H				1.87
D35	K	CH ₂ NHCH ₃				1.61
D36	K	CH ₂ OH				1.68
D37	L	CH ₂ NHCH ₃				1.68
D38	L	CH ₂ OH				2.61
D39	M					2.69
D40	N					2.37
D41	O	H				4.11
D42	O	CH ₂ NHCH ₃				2.65
D43	P	CH ₂ NHCH ₃				2.33
D44	P	CH ₂ OH				3.94

Table 1(e). Eighty molecules obtained using the CoMFA/LINK approach.

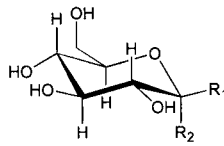
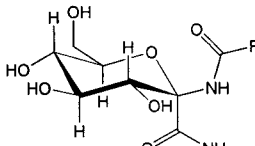
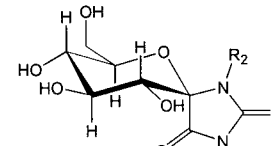
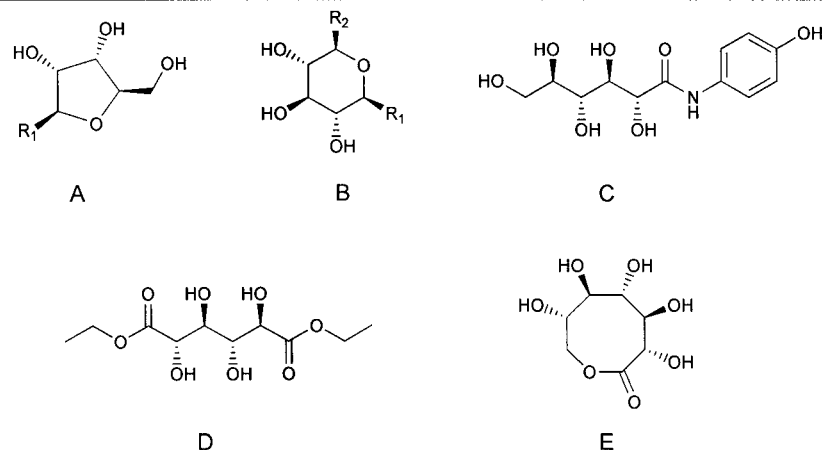
<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">  <p>A (GP1)</p> </div> <div style="text-align: center;">  <p>B (GP2)</p> </div> <div style="text-align: center;">  <p>C (GP3)</p> </div> </div>									
no.	str.	R ₁	R ₂	pK _{i,final}	no.	str.	R ₁	R ₂	pK _{i,final}
E1	A	CH ₂ CH=CH ₂	CF ₂ H	2.63	E41	C	CF ₂ H	CH ₂ OH	2.19
E2	A	CH ₂ Cl	CN	2.82	E42	C	CF ₂ H	CF ₂ H	2.31
E3	A	CH ₂ OCH ₃	CF ₂ H	3.07	E43	C	CF ₂ H	H	2.96
E4	A	CH ₂ OCH ₃	CN	2.90	E44	C	CN	CH ₂ Cl	2.28
E5	A	CH ₂ SO ₂ CH ₃	CF ₂ H	2.42	E45	C	CN	CH ₂ F	2.23
E6	A	CH ₂ SO ₂ CH ₃	Cl	2.37	E46	C	CN	CH ₂ OH	2.22
E7	A	CH ₂ SO ₂ CH ₃	Br	2.19	E47	C	CN	CF ₂ H	2.33
E8	A	NHCH ₂ CH ₃	CN	3.46	E48	C	CN	Cl	2.27
E9	A	NHCOCH ₃	Cl	3.66	E49	C	H	CH ₂ Cl	3.63
E10	A	NHCOCH ₃	CN	3.44	E50	C	H	CH ₂ F	3.81
E11	A	SCONH ₂	Cl	3.16	E51	C	H	CH ₂ OH	4.09
E12	A	SCONH ₂	CN	2.91	E52	C	OH	CH ₂ F	4.04
E13	A	SCONH ₂	CONH ₂	3.13	E53	C	SOCH ₃	CH ₂ Cl	1.99
E14	A	SCONH ₂	CO ₂ H	3.06	E54	C	SOCH ₃	CH ₂ F	2.35
E15	A	SCONH ₂	Br	3.19	E55	C	SOCH ₃	CH ₂ OH	2.02
E16	B	C ₆ H ₅		2.91	E56	C	SOCH ₃	CF ₂ H	2.22
E17	B	CH ₂ CN		4.36	E57	C	SCONH ₂	CH ₂ Cl	1.82
E18	B	CH ₂ F		4.30	E58	C	SCONH ₂	CH ₂ F	1.91
E19	B	CH ₂ NO ₂		2.34	E59	C	SCONH ₂	CH ₂ OH	1.85
E20	B	CH ₂ OH		4.23	E60	C	SCONH ₂	CF ₂ H	1.90
E21	B	CHO		4.13	E61	C	SCONH ₂	Cl	1.79
E22	B	CHCl ₂		3.23	E62	C	SCONH ₂	CO ₂ H	1.84
E23	B	NHSO ₂ NH ₂		3.00	E63	C	SCONH ₂	H	2.41
E24	B	C ₆ H ₄ Cl		2.99	E64	C	SCONH ₂	NH ₂	1.84
E25	B	NHOH		4.54	E65	C	SCONH ₂	OCH ₃	1.67
E26	B	NO		4.26	E66	C	SCONH ₂	OH	1.89
E27	B	SCOCF ₃		3.95	E67	C	SCONH ₂	Br	1.80
E28	B	SCN		4.45	E68	C	SO ₂ CF ₃	CH ₂ Cl	1.78
E29	B	Tetrazol-1-yl		3.22	E69	C	SO ₂ CF ₃	CH ₂ F	1.82
E30	B	C(CH ₃)=CH ₂		2.74	E70	C	SO ₂ CF ₃	CH ₂ OH	1.78
E31	C	CH ₂ Cl	CH ₂ Cl	1.95	E71	C	SO ₂ CF ₃	CF ₂ H	2.12
E32	C	CH ₂ Cl	CH ₂ F	1.85	E72	C	SO ₂ CF ₃	Cl	1.71
E33	C	CH ₂ F	CH ₂ Cl	2.29	E73	C	SO ₂ CF ₃	CONH ₂	1.86
E34	C	CH ₂ F	CH ₂ F	2.52	E74	C	SO ₂ CF ₃	H	1.85
E35	C	CH ₂ F	CH ₂ OH	2.54	E75	C	SO ₂ CF ₃	NH ₂	1.74
E36	C	CH ₂ F	Cl	2.39	E76	C	SO ₂ CF ₃	OCH ₃	1.70
E37	C	CH ₂ OH	CH ₂ F	2.96	E77	C	SO ₂ CF ₃	OH	1.72
E38	C	CH ₂ OH	CH ₂ OH	3.16	E78	C	SO ₂ CH ₃	CH ₂ Cl	1.86
E39	C	CF ₂ H	CH ₂ Cl	2.06	E79	C	SO ₂ CH ₃	CH ₂ F	1.98
E40	C	CF ₂ H	CH ₂ F	2.23	E80	C	SO ₂ CH ₃	CH ₂ OH	1.79

Table 1(f). Ten hits from the HQSAR/NCI database search.



no.	structure	R ₁	R ₂	pK _{i,final}
F1	A	NHCONHOH		2.23
F2	B	NHCONHCH ₂ CH ₂ Cl	CH ₂ OH	3.23
F3	C			2.91
F4	B	NHCOCH ₂ Br	CH ₂ OH	3.76
F5	A	NHCONHCH ₃		1.89
F6	A	CONH ₂		2.08
F7	D			1.53
F8	B	=O	CH ₂ OH	2.99
F9	E			2.07
F10	B	NHCONHCH ₂ CH ₂ Cl	H	2.41

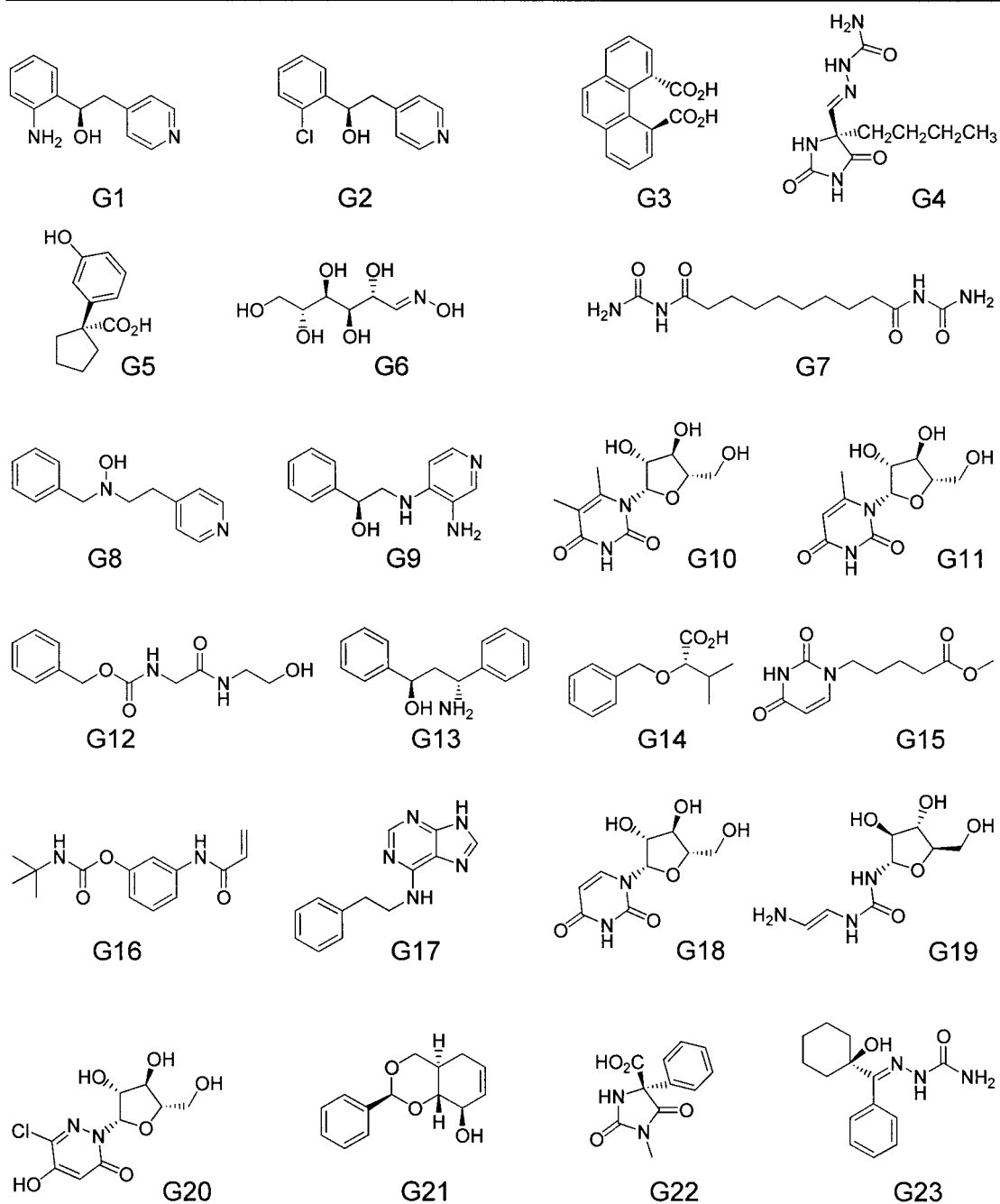
A version of the available chemical directory (ACD) database, which contains a total of 65872 structures, was screened using the LUDI program against GP. The residues inside a 10 Å radius sphere, which centered on the C1 atom of the glucose molecule, was used to generate the interaction sites. Other settings were kept at the default values. A total of 410 interaction sites, including 153 HB acceptors, 220 HB donors and 37 lipophilic centers, were identified in the binding pocket. The search of the ACD took approximately 20 CPU h on a 175 MHz R10000 Silicon Graphics Octane workstation. In all, 64 candidates were selected from the database with a minimum LUDI score of 300, the cutoff value in the preliminary filter layer in the prediction system. The top ten candidates, **C1-10** (Table 1(c)), whose scores ranging from 508 to 390, were selected for further analysis. These molecules are generally small in size, and may be used as scaffolds for linking other types of fragments. Fig-

ure 7 shows the positions of the best ligands in the bind site, where each of them is making about five hydrogen bonds with the protein. An interesting result that hydantoin (substructure A in Table 1(c)) appears to be a promising lead (**C2** in Figure 7). That same substructure is present in the most active GP ligand to date (**GP3**), though experimentally the substructure is found to bind at a different location since the site found by LUDI is occupied by the glucose moiety (Figure 8).

(d) LUDI/LINK

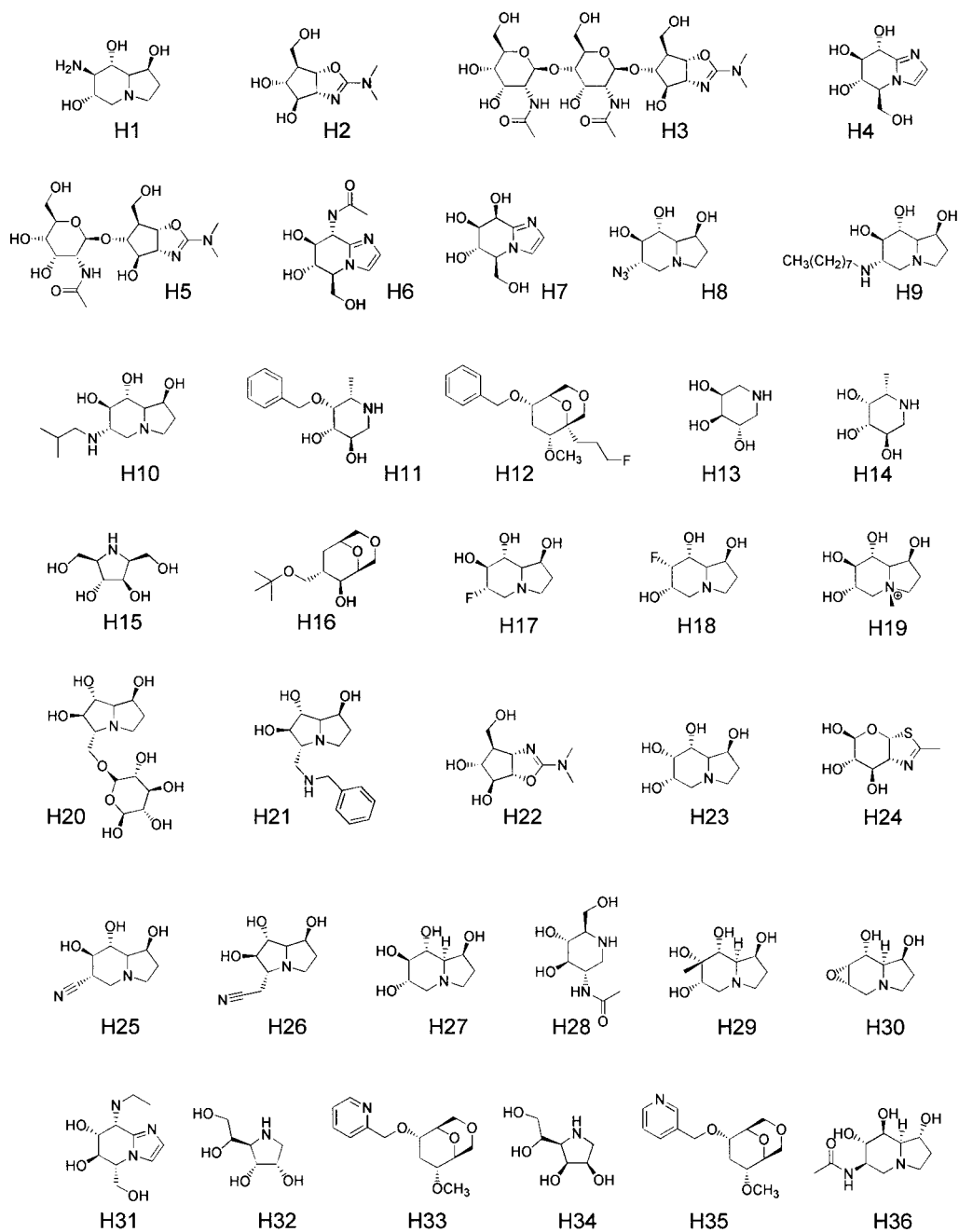
The link mode of the LUDI program was invoked to generate *de novo* ligands by connecting molecular fragments from its standard link library in LUDI (comprised of approximately 1100 entries) to a number of template molecules. LUDI used the same set of geometric rules to fit the fragments into the binding site; and in addition, it determined whether these fragments can be connected to the template with a

Table 1(g). Twenty-three hits obtained from the SMGNN/NCI database search.



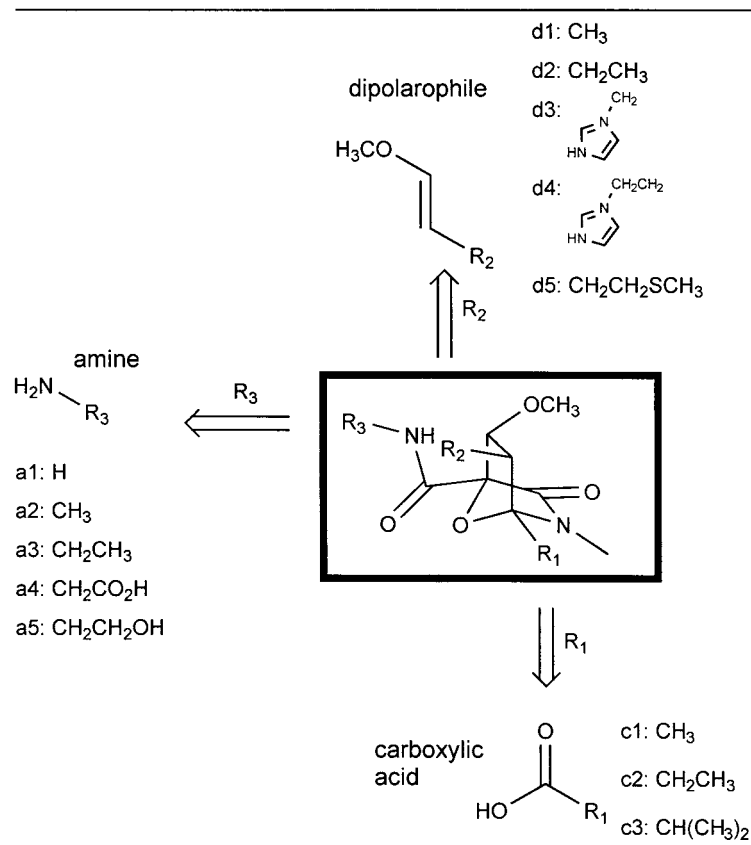
no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$
G1	1.77	G5	1.73	G9	1.55	G13	1.78	G17	1.85	G21	1.77
G2	1.69	G6	2.57	G10	1.87	G14	1.79	G18	2.40	G22	1.80
G3	1.52	G7	1.50	G11	1.99	G15	2.14	G19	2.66	G23	1.38
G4	1.58	G8	1.81	G12	1.95	G16	1.58	G20	3.16		

Table 1(h). Thirty-six compounds in the SDB.



no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$
H1	1.16	H7	2.53	H13	1.78	H19	1.32	H25	1.38	H31	1.76
H2	2.06	H8	1.35	H14	1.88	H20	1.89	H26	1.39	H32	2.29
H3	0.69	H9	0.89	H15	2.21	H21	1.37	H27	1.22	H33	1.50
H4	1.58	H10	1.31	H16	0.94	H22	1.71	H28	1.48	H34	1.80
H5	0.81	H11	1.65	H17	1.19	H23	1.34	H29	1.35	H35	1.72
H6	1.42	H12	1.25	H18	1.20	H24	1.78	H30	1.31	H36	1.73

Table 1(i). Seventy-five CombiChem products.



reasonable bond geometry. In this method, three existing GP ligands (**GP1-3**), together with the ten LUDI/ACD candidates (**C1-10**) found in the previous section, served as the templates. Any hydrogen atoms in the template structure could be replaced by a link fragment to create a new ligand. More than one fragments might be attached to the same template molecule, provided that there were no unfavorable intramolecular contacts between the fragments. Figure 9 shows how two fragments, which contain functional groups that are consistent with the LUDI interaction sites, can be simultaneously linked to **GP2** to form **D42** (Table 1(d)).

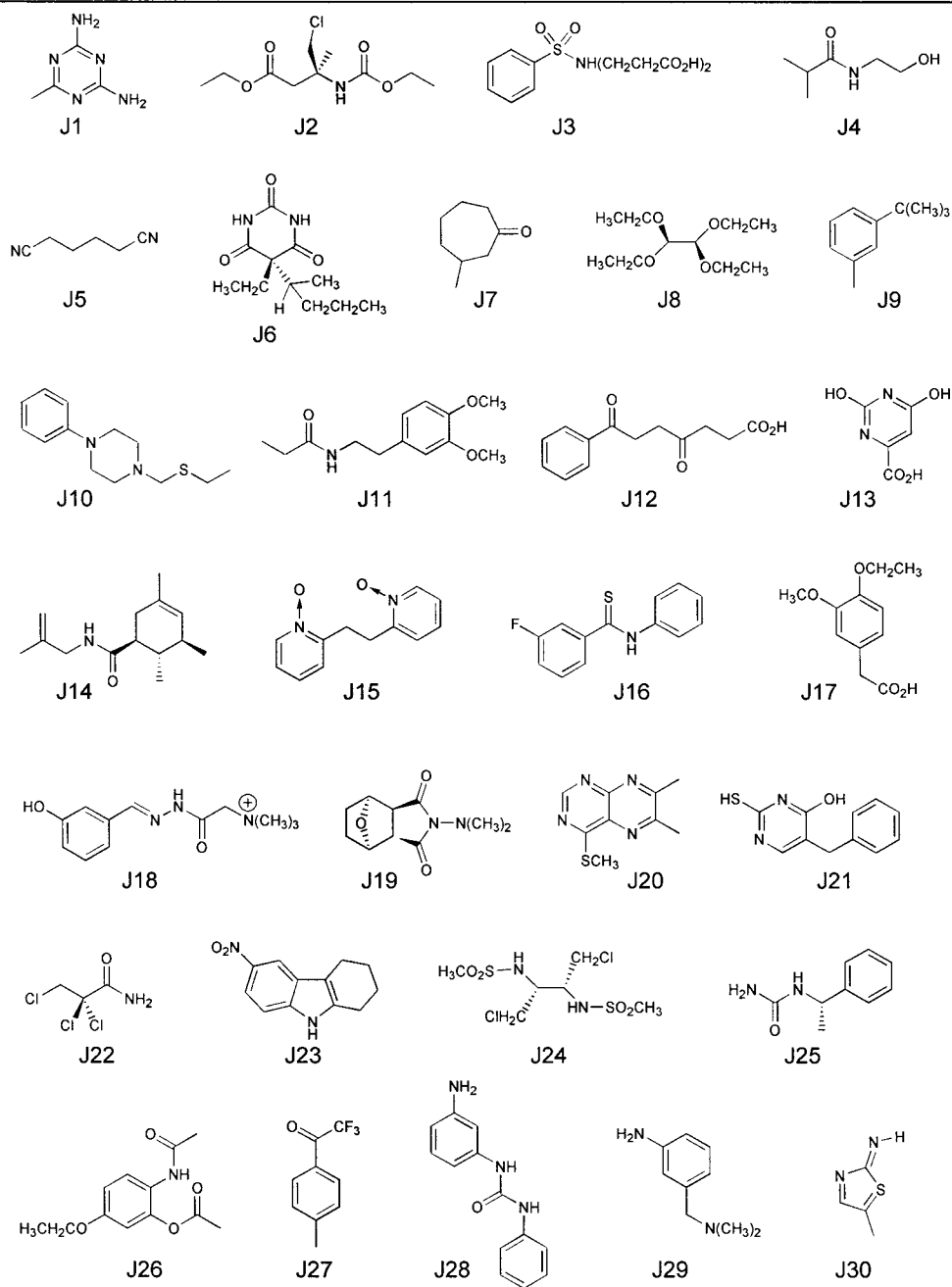
Thirty-five link fragments, which made complementary interactions with the protein, were identified by the LUDI program as suitable extensions to the existing template molecules. A total of 44 new compounds were generated by connecting one or more fragments to the templates. Ten of those compounds

(**D1-4** and **D39-44**) were glucose derivatives; the remaining (**D5-38**) were based on the earlier LUDI/ACD candidates. The chemical structures are shown in Table 1(d).

(e) CoMFA/LINK

A CoMFA model based on 30 GP ligands with known structures was reported in the previous paper [32]. The model, which employed AM1 charges, was a four-component PLS QSAR that yielded r^2 and q^2 values of 0.93 and 0.60, respectively. We utilized this QSAR model in conjunction with the advanced CoMFA module in SYBYL to construct potent ligands. In this method, different substituents were attached to the pre-defined positions of an existing template to form new candidate ligands. For each trial ligand, a SYBYL programming language script was executed to obtain AM1 charges, and the CoMFA QSAR was used to predict its biological activity. Furthermore, this pro-

Table 1(j). Thirty compounds chosen at random from the NCI database.



no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$	no.	$pK_{i,final}$
J1	2.68	J6	2.13	J11	0.86	J16	1.88	J21	1.77	J26	1.63
J2	1.82	J7	1.37	J12	1.72	J17	1.59	J22	3.51	J27	1.44
J3	1.75	J8	1.32	J13	2.72	J18	1.26	J23	1.63	J28	1.60
J4	1.68	J9	1.77	J14	1.35	J19	1.42	J24	1.60	J29	1.59
J5	2.25	J10	1.57	J15	1.32	J20	1.49	J25	1.79	J30	1.80

Table 2. The MCSS minima used by the OLIGO program for peptide constructions.

Structural component	MCSS group	Number of minima	Reference energy <i>in vacuo</i>	Energy range min, max
Main-chain	NMAC	112	−3.49	−39.25, −0.29
Asn side-chain	ASNR	191	−12.44	−38.24, 2.52
Gln side-chain	GLNR	278	−11.61	−39.93, 2.63
Ser side-chain	MEOH	128	0.00	−30.74, 1.24
Thr side-chain	THRR	232	−0.07	−30.90, 2.99

gram also examined the contribution to activity that was derived from molecular regions where there was only limited chemical information in the training data, thereby giving an estimate of the amount of data extrapolation involved in the prediction process. Numerically, the extrapolation for a given compound k is obtained by summing the contributions from each of the field descriptors (q) that has gone out-of-range relative to the extreme values defined by the training compounds:

$$\text{Extrapolation}_k = \sum_q b_q \times (X_{kq} - \overline{X}_q), \quad (1)$$

where b_q is the coefficient of the descriptor in the linear PLS regression, X_{kq} is the field value for k and \overline{X}_q is its average value over the training set of compounds.

The three standard GP ligands (**GP1-3**) were again employed as the molecular templates, and the substituent attachment positions for each molecule are highlighted in Table 1(e). For **GP1**, the β - and α -substituent positions at C1 (R_1 and R_2 of template **A** in Table 1(e)) were the suitable attachment points because most of the data variance of the training compounds was derived from this region. A total of 476 new analogs were generated by all possible substituent combinations from two lists containing 34 (R_1) and 14 (R_2) functional groups. **GP2** has only one variable position that connects the carbonyl carbon of its β -substituent, thereby allowing a larger list of 246 substituent groups to be screened. The two nitrogen atoms of **GP3** were used as attachment points, and similar to **GP1**, 476 analogs were tested by the systematic addition of pairs of the same 34 and 14 functional groups to each of the two positions.

The entries **E1-15** listed in Table 1(e) are the 15 derivatives of **GP1** that were predicted to bind

best by CoMFA. Their predicted $\text{pK}_{i,\text{CoMFA}}$ values, based on CoMFA calculations, ranged from 4.44 to 3.90. Because the contribution from extrapolations was low (an average of 0.26 for the 15 compounds), these predictions should be reasonably reliable. The top 15 derivatives of **GP2** (**E16-30**) have predicted $\text{pK}_{i,\text{CoMFA}}$ values from 5.36 to 5.06. All of them have considerably higher predicted values than the parent compound ($R_1 = \text{OCH}_3$; $\text{pK}_{i,\text{CoMFA}} = 4.83$) although the average contribution from extrapolation (0.59) is somewhat larger than for the **GP1** derivatives. The derivatives of **GP3** displayed the highest predicted activity with the CoMFA model. The predicted $\text{pK}_{i,\text{CoMFA}}$ values of the best 50 candidates ranged from 6.07 to 5.21, well above that of the parent compound ($R_1 = R_2 = \text{H}$; $\text{pK}_{i,\text{CoMFA}} = 4.88$). However, most of these compounds, particularly those with the highest predicted values, involve extrapolation beyond 1.0, the recommended threshold of reliable prediction. For this reason, we decided to examine all the 50 **GP3** derivatives (**E31-80**) using the hybrid prediction system.

(f) HQSAR/NCI

In the previous publication [32] a hologram QSAR (HQSAR) model was described for the 30 GP inhibitors. This QSAR model included hydrogen atoms and chirality information in fragment generation, and yielded an r^2 value of 0.88 and a q^2 of 0.65. We used this model to search for novel leads from the National Cancer Institute (NCI) library, which is comprised of approximately 127,000 compounds. Tripos Inc. supplied a 2D version of this library in UNITY

It is important to distinguish between $\text{pK}_{i,\text{CoMFA}}$ and the $\text{pK}_{i,\text{final}}$ values reported in Table 1. The former refers to the predicted pK_i value by CoMFA during the ligand design stage, whereas the latter was derived from all QSAR and SBEP prediction methods. Thus, the two values can be quite different. Likewise, $\text{pK}_{i,\text{SMGNN}}$ and $\text{pK}_{i,\text{HQSAR}}$ refer to the original activity predictions by SMGNN and HQSAR programs during the ligand design stage.

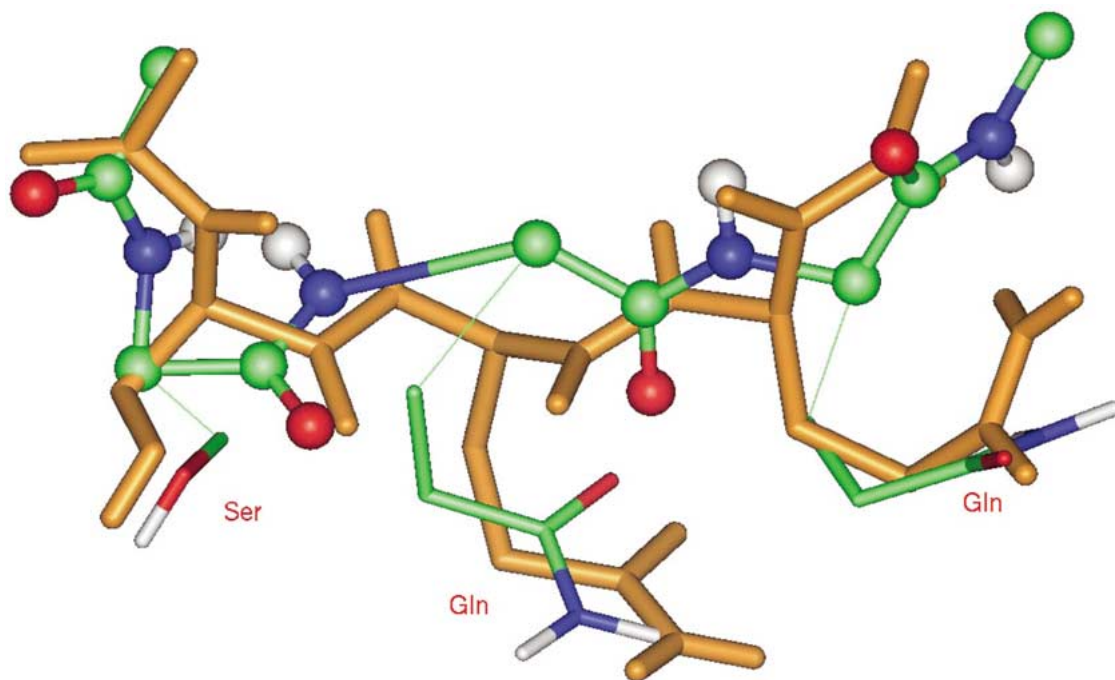


Figure 6. Construction of a tri-peptide **B4** (D-Ser-D-Gln-D-Gln) from MCSS functional group minima using the OLIGO program. Four *N*-methylacetamide (NMAC) MCSS minima are connected to form a peptide backbone, and a Ser and two Gln MCSS minima are attached to the C $_{\alpha}$ atoms along the peptide chain. The conformation of the new peptide is then regularized by energy minimization in the fixed protein, and the final structure is shown in orange.

database format, which could be accessed directly by the HQSAR module. Because of its large volume, a preliminary filter was established so that only entries exceeding a 2D similarity (Tanimoto coefficient) [46] cutoff of 0.85 with any of the 30 known GP ligands would be examined.

A total of 107 hits were found in the database with predicted $\text{pK}_{i,\text{HQSAR}}$ values¹ ranging from 3.99 to -8.56 . The top ten candidates (Table 1(f): **F1-10**), having predicted activities between 3.99 and 2.98, were selected for further analysis. It should be noted that the candidates suggested by the HQSAR program do not contain any spatial reference to the GP molecule. To generate spatial coordinates that would be necessary for use in other 3D QSAR predictors, the 2D molecular representation was converted to 3D coordinates using SYBYL. The initial alignment was obtained using a fieldfit algorithm available in the Search_Compare module of the INSIGHT program. Specifically, the most active compound (**GP3**) was employed as a template for flexible fitting, where the orientation or conformation of the ligand may be modified in such a way that its similarity (calculated from

molecular fields) with the template molecule would be maximized.

(g) SMGNN/NCI

In the previous paper [32] we reported a predictive ($q^2 = 0.82$) neural network QSAR model based on six molecular similarity descriptors. In this study, we used this SMGNN model to perform a 3D database search on the NCI library of compounds. The 3D form of NCI database was obtained from the NCI public ftp server, where the 2D to 3D structure conversion was performed by the Corina program [47]. A preliminary filter was introduced to examine compounds containing up to 20 non-hydrogen atoms that were among the most common element types in organic chemistry; i.e., C, N, O, F, P, S, Cl, Br, and I. Furthermore, since the distributed library did not include hydrogen, the INSIGHT program [48] was used to add the appropriate hydrogen atoms. The alignment of the candidate ligands was made using the fieldfit procedure described in (f). The six similarity descriptors used in the SMGNN QSAR model were obtained from the Search_Compare module of INSIGHT. Finally, the values of the descriptors were passed through the

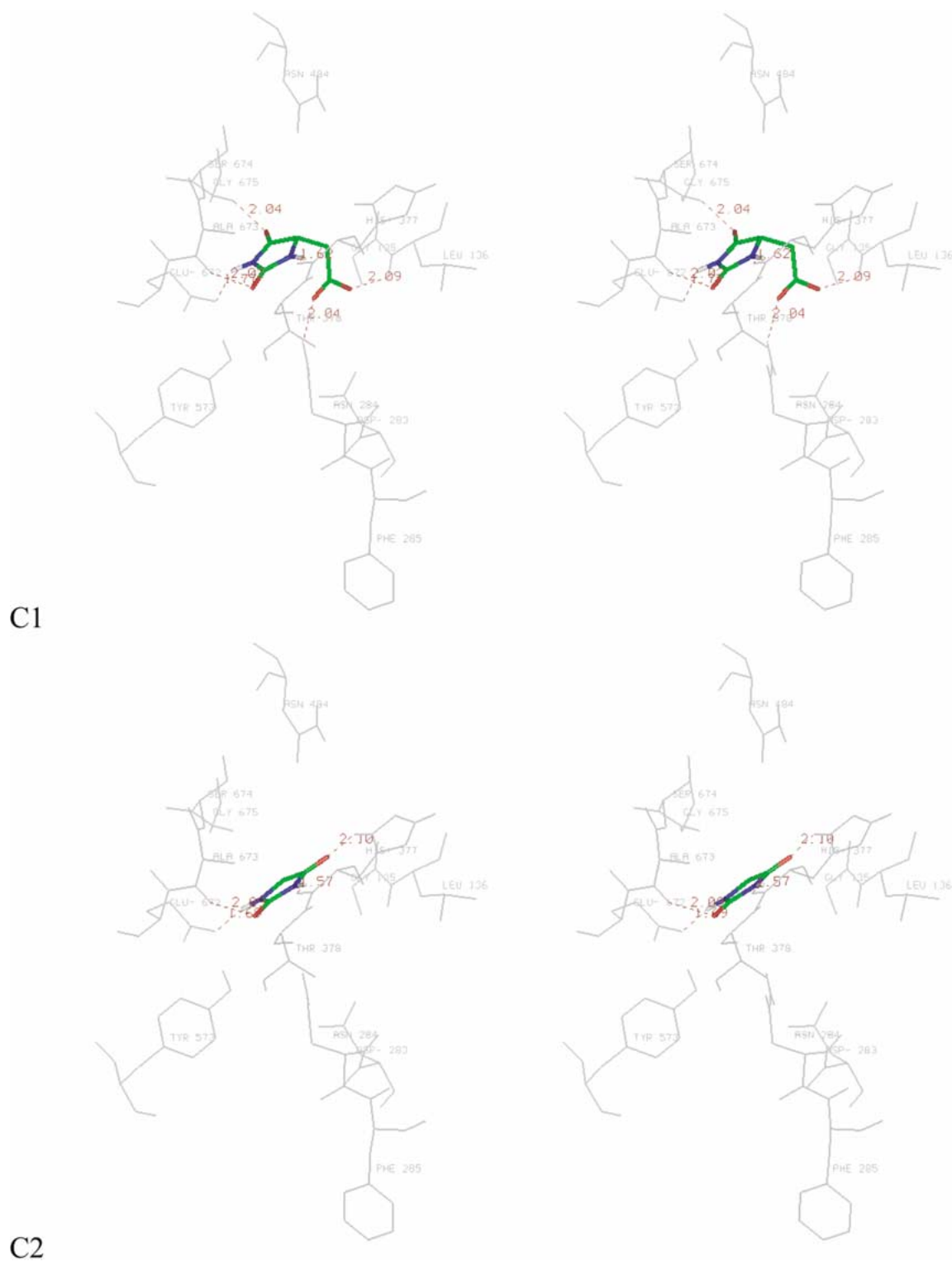


Figure 7. The top four hits (**C1-4**) found by the LUDI program from the ACD. The intermolecular hydrogen bonds between the four molecules and the binding sites are dotted.

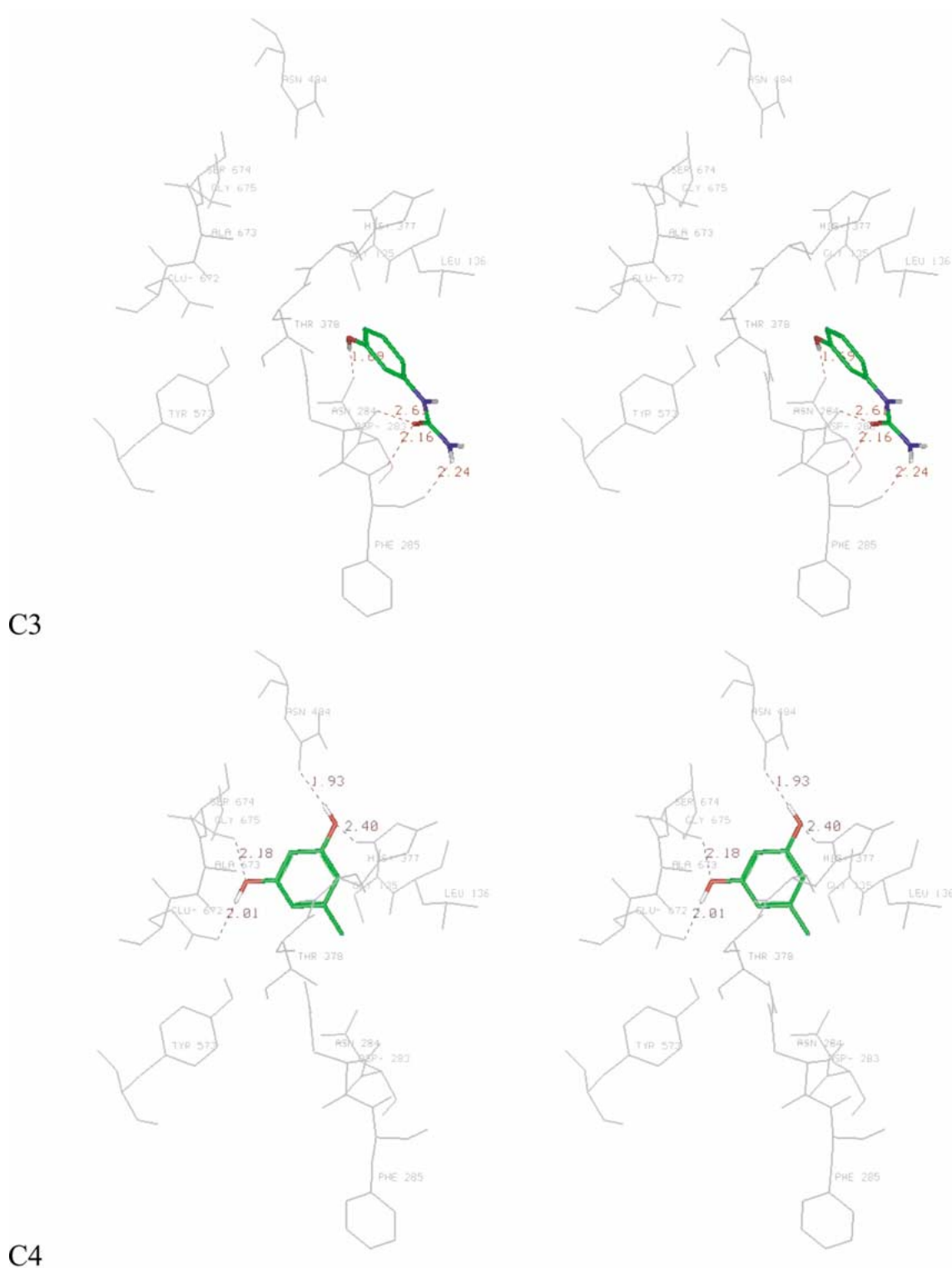


Figure 7. Continued.

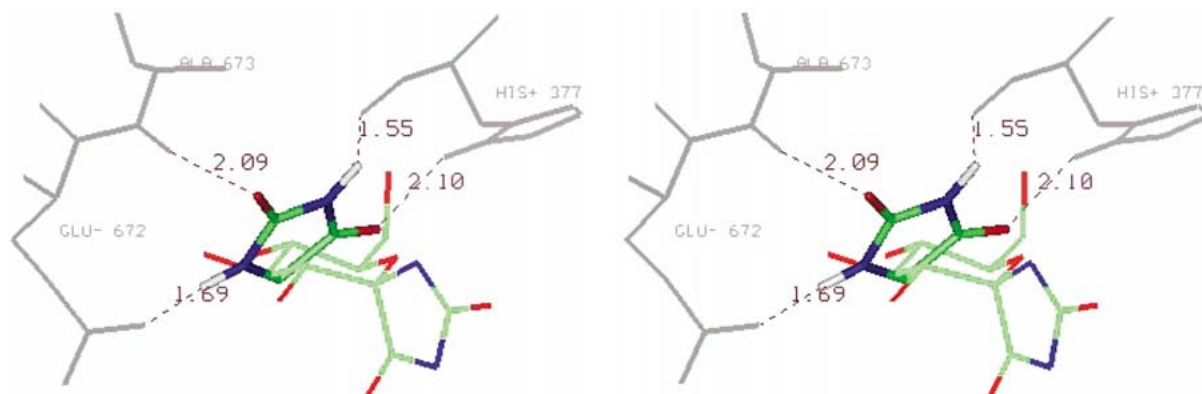


Figure 8. A stereoview of the hydantoin (**C1**) identified by LUDI/ACD. All four N-H and C=O functional groups in this molecule form intermolecular hydrogen bonds with the GP binding site, and it is predicted to bind in a different location than the hydantoin substructure of **GP3**.

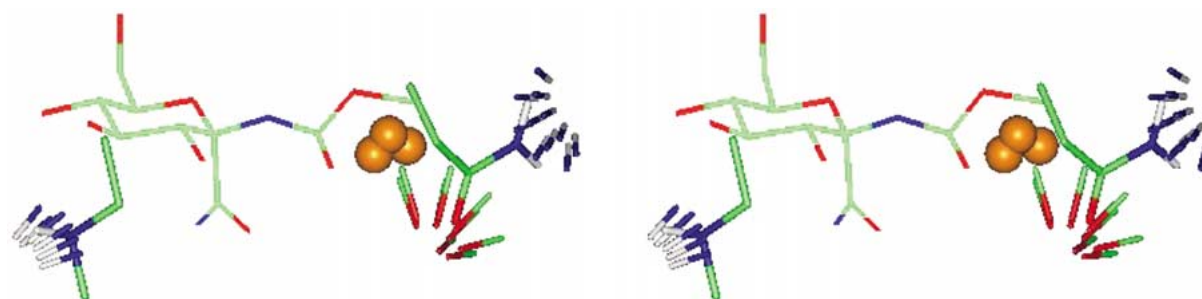


Figure 9. Linking two LUDI fragments to **GP2** to form **D42**. The two fragments possess functionalities that are consistent with the LUDI interaction map and are in favorable orientation for direct attachment to the lead molecule.

neural network and the predicted pK_i values of the candidate ligands were obtained.

The predicted $pK_{i,SMGNN}$ values¹ of the top 23 ligands (Table 1(g): **G1-23**) ranged from 5.24 to 4.40, and are comparable to those for the most potent compounds (**GP2**: $pK_{i,SMGNN} = 4.87$; **GP3**: $pK_{i,SMGNN} = 4.88$). Although this set of compounds is chemically diverse, all of them have multifunctional groups that form many hydrogen bonds with the protein.

(h) FieldFit/SDB

Prof. L.N. Johnson and co-workers at Oxford University provided us with a small database (SDB) of chemical structures that they had planned to screen. This database contained 36 poly-hydroxyl compounds (Table 1(h): **H1-36**), which have certain chemical resemblance to glucosides. Several modules of the INSIGHT program were used to construct the 3D coordinates for the compounds from their chemical structures. First, the chemical connectivity of the molecules was established using the molecular sketcher facility.

The 2D sketches were processed using the Converter program to yield initial 3D coordinates. Using the FieldFit algorithm of the Search_Compare module, the ligands were docked to the binding site of GP with the binding conformation of **GP3** serving as a reference structure.

(i) FieldFit/CombiChem

This study makes use of a combinatorial library system developed in Prof. D.J. Austin's laboratory at Yale University. A retrosynthetic analysis, shown in Table 1(i), outlines how each monomeric segment is assembled to form the scaffold (highlighted in the center). In our current analysis we consider three carboxylic acid monomers ($R_1 = c1-3$), five dipolarophile monomers ($R_2 = d1-5$) and five amine monomers ($R_3 = a1-5$). This set of monomers yielded a total of 75 compounds in this virtual library (**I1-75**). The carboxylic acids and amine monomer units are commercially available, while the dipolarophiles may require prior synthesis.

A new MCSS group was created based on the chemical structure of **II**, the smallest combinatorial product in the library. Its 3D structure was modeled based on the x-ray coordinates of an analogous compound. A functionality map of this group was obtained using the MCSS procedure described above. 128 minima were found in the glucose-binding region. The second best MCSS minimum (sorted according to interaction energy) formed good hydrogen bonds with the protein, and the substituent groups (R₁-R₃) pointed to empty binding pockets so that the more bulky analogs could be accommodated. We took this conformation as the template, and used the Analog Builder of Cerius² [49] to build a virtual combinatorial library of 75 compounds by connecting the appropriate substituent groups corresponding to each of the monomers to the basic scaffold.

(j) Random/NCI

A control calculation was performed to assess the effectiveness of the design strategies described above. Thirty compounds (Table 1(j)) were taken at random from the NCI library and were fitted into the binding site of GP using the method described in (h). Their predicted binding affinities against GP can be regarded as a reference point for other designed compounds.

III. Structure refinement

Some of the above methods are based on existing GP ligands and do not implicitly include information about the binding site. Thus, it is possible that designed ligands have unfavorable steric contacts with the enzyme. To obtain a more realistic binding conformation, the structure of the candidate ligand was energy minimized in the binding site. The DISCOVER program and the CFF91 parameters were used for the structure refinement. Specifically, the ligands were minimized inside a fixed protein (with an all hydrogen representation) using a two-stage minimization protocol. In the first stage (250 steps) the heavy atoms of the ligands were harmonically restrained at the original positions and only the hydrogen atoms were free to move. In the second stage (500 steps), the restraints were removed and all ligand atoms could move. A minimization algorithm, which switched from steepest descent to conjugated gradient method when the gradient was less than 1000 kcal/mol Å, was used. The minimization stopped when a final rms gradient of 0.001 kcal/mol Å was reached. For the evaluation of the non-bonded interactions, a distance dependent

Table 3. The number of the designed and random ligands that are removed by each of the layers in the hybrid prediction system.

	Designed set	Random set
Total no. of ligands	301	30
No. of ligands removed by filter layer	19	10
No. of ligands removed by prediction layer	163	12
No. of ligands removed by validation layer	70	8
Final no. of candidate ligands	49	0

dielectric of 1.0 and an atom-based cutoff of 12 Å were employed. The energy minimized conformations for the ligands were used for the final activity prediction phase. Due to this refinement process, the final pK_i predictions for a given ligand may be different from the values obtained by the QSAR or LUDI calculations during the design phase.

Analysis

The ligand design procedures outline in the Method sections led to the construction of 301 candidate ligands. These ligands, together with the 30 compounds that had been chosen at random from the NCI database, were evaluated by the three-stage scoring scheme described in Methods and Figure 2. Forty-nine out of 301 designed compounds were established as the leading candidates for potential GP ligands. As expected, none of the 30 randomly chosen compounds remained after the three stages of evaluation. Table 3 shows the number of compounds that were removed by each of the layers in the scoring scheme. The LUDI system appeared to be an effective filter, particularly for discriminating against the randomly chosen compounds. Only 6% of the designed set was removed at this stage compared with 33% of the random set. Approximately half of the designed ligands were discarded in the prediction phase, and another quarter were removed at the validation stage. The remaining 49 candidates are listed in Table 4(a), together with their respective scores at the three stages.

Computational efficiency was an important consideration in the development of the current scoring function. A major assumption was that a significant fraction of the compounds would be eliminated in the preliminary and the prediction layers so that only a small number of candidate molecules would require

Table 4(a). Forty-nine primary candidates identified by the hybrid prediction system. Forty-four of them are glucose analogs. The predicted pK_i values are derived from the QSAR and SBEP predictions according to Equations 2 and 3. The four non-glucose derivatives are shown in bold typeface.

Rank	Compound	LUDI score	Average QSAR	SBEP	Predicted pK_i
R1	E25	767	3.51	5.17	4.54
R2	E28	639	3.68	4.85	4.45
R3	E17	626	3.18	5.19	4.36
R4	E18	669	3.60	4.68	4.30
R5	E26	564	3.27	4.92	4.26
R6	E20	667	3.45	4.71	4.23
R7	A8	643	3.19	4.88	4.19
R8	E21	611	3.54	4.45	4.13
R9	D41	692	2.99	4.95	4.11
R10	E51	617	3.43	4.47	4.09
R11	E52	865	3.37	4.45	4.04
R12	E27	693	2.71	4.94	3.95
R13	D44	606	3.23	4.41	3.94
R14	A7	580	3.69	3.88	3.89
R15	A9	740	3.29	4.15	3.82
R16	E50	680	3.55	3.88	3.81
R17	A1	793	3.54	3.82	3.77
R18	F4	773	3.04	4.30	3.76
R19	E9	788	2.99	4.18	3.66
R20	E49	666	2.94	4.17	3.63
R21	E8	751	3.11	3.71	3.46
R22	E10	746	2.98	3.81	3.44
R23	D4	550	2.60	4.16	3.44
R24	F2	673	2.97	3.44	3.23
R25	E22	683	2.81	3.60	3.23
R26	A6	524	3.21	3.21	3.22
R27	E29	590	2.97	3.44	3.22
R28	E15	682	2.73	3.61	3.19
R29	E11	676	2.79	3.51	3.16
R30	G20	663	3.02	3.28	3.16
R31	E38	701	3.03	3.26	3.16
R32	E13	611	2.69	3.55	3.13
R33	E3	639	3.11	3.05	3.07
R34	E14	538	2.55	3.57	3.06
R35	A3	869	3.22	2.91	3.06
R36	D1	610	2.52	3.56	3.04
R37	E23	678	2.89	3.14	3.00
R38	E24	842	3.07	2.95	2.99
R39	F8	434	2.79	3.21	2.99
R40	E37	773	3.41	2.55	2.96
R41	F3	577	2.63	3.23	2.91
R42	E16	697	2.59	3.27	2.91
R43	E12	559	2.60	3.25	2.91
R44	E4	590	3.20	2.65	2.90
R45	E2	598	2.81	2.90	2.82
R46	E30	664	3.08	2.50	2.74
R47	D5	324	2.57	2.97	2.72
R48	D6	381	2.84	2.59	2.66
R49	E1	645	2.56	2.82	2.63

the CPU intensive calculations in the final stage. In this study, approximately one-third of the compounds reached the validation stage. If such percentage is rep-

resentative for a typical set of designed compounds in a larger sample, the scoring procedure can handle on the orders of 10^4 ligands on a 16-processor R10000

Table 4(b). Twelve secondary candidates identified by the hybrid prediction system. Only non-glucose derivatives are considered in this analysis. The predicted pK_i values are derived from the QSAR and SBEP predictions according to Equations 2 and 3.

Rank	Compound	LUDI score	Average QSAR	SBEP	Predicted pK_i
R50	C4	353	2.23	4.49	3.41
R51	D30	537	2.06	4.52	3.33
R52	C1	265	2.99	2.92	2.93
R53	D8	453	2.39	3.25	2.78
R54	D7	262	2.59	3.02	2.76
R55	G19	380	3.04	2.41	2.66
R56	D38	292	2.69	2.67	2.61
R57	H7	447	2.29	2.91	2.53
R58	D29	388	2.68	2.42	2.46
R59	G18	527	2.99	2.01	2.40
R60	H32	402	2.22	2.57	2.29
R61	H15	437	2.00	2.64	2.21

SGI workstation in about 2 weeks, which could be acceptable for practical purposes. Over 90% of the CPU time in the validation was spent in the evaluation of ΔG_{elec} . If one can optimize the procedure to calculate (or use alternative descriptors to replace) this term, a much faster turnover time can be realized. One possibility is the analytical continuum electrostatics (ACE) method [50], which provides an analytical treatment of the continuum model for electrostatic solvation and is more than two orders of magnitude faster than numerical finite-difference continuum calculations that were used in this study.

Because the activity prediction system has been parameterized using glucose analogs, there is a strong bias favoring this chemical class in the prediction. In fact, all but 4 of the 49 candidates emerged from the initial analysis are glucose derivatives; the four are shown in bold typeface in Table 4(a). In light of this, it is appropriate to consider candidates of other chemical classes that marginally failed to meet the selection criteria. Since the total number of designed compounds constructed in this study was relatively small, it was feasible to perform the full QSAR and SBEP predictions on the entire set. A total of 12 second tier candidates, of which none are unrelated to glucose, were identified and they are listed in Table 4(b).

The predicted activities for the primary compounds ($pK_{i,\text{final}}$: 4.54 to 2.63) and secondary compounds (3.41 to 2.21) are within the predicted range corresponding to the 30 inhibitors with known x-ray structures (4.61 to 1.48). On an individual basis, the predicted $pK_{i,\text{final}}$ of the best candidates are compara-

ble to the most potent compounds (**GP2**: 4.61; **GP3**: 4.50). The majority of the 301 ligands are predicted to bind to the glucose-binding site in the protein. The binding conformations of the glucose derivatives are similar to their parent structures, in that they maintain the essential hydrogen bonding network with the enzyme. The few molecules bound elsewhere are the novel entities constructed by LUDI or OLIGO. One of them (**C7**) is founded almost 9 Å away from the main site, as measured by the distance from its center of mass (CM) to that of **GP3**. Figure 10 shows a scatter plot of the score of the ligand against its CM distance to the major binding region (defined by the CM of **GP3**). It seems that, due to the QSAR-based component of the scoring function, binding in the vicinity of the glucose-binding site was an important, though not sufficient, criterion for a ligand to achieve high ranking. None of the primary and only one (**C8**, labeled in the figure) of the secondary candidates occupied a site more than 2.5 Å away from the main site.

To rationalize the important factors that determine the predicted binding affinity, the binding site is separated into two main binding pockets, which are further divided into five sub-regions. The first binding pocket consists of three sub-regions which contain the glucose moiety: the pyridoxal 5'-phosphate (PLP) cofactor, the hydrophilic ring of residues forming the majority of hydrogen bonds with the hydroxyl groups of glucose, and the hydrophobic residues that are mostly in contact with the non-polar glucose atoms. The second binding pocket interacts with the glucose C1 substituents. The α - and β -regions, which are asso-

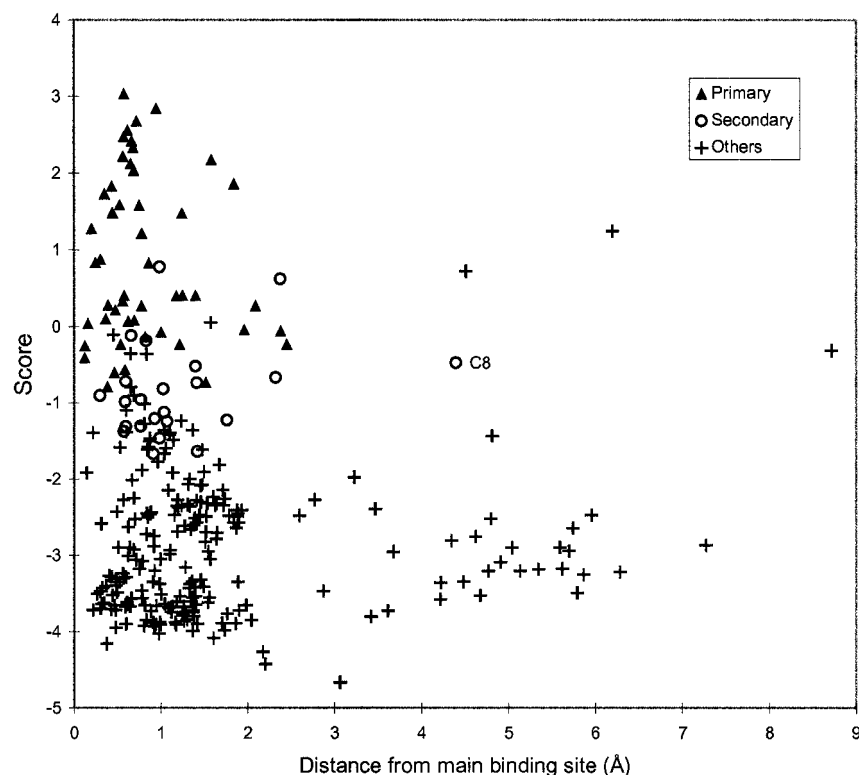


Figure 10. A scatter plot of the final score of a ligand against its distance to the glucose-binding site. The distance is defined by the center of mass of the ligand to that of **GP3**.

ciated with the stereochemical orientation of the two substituents, belong to this pocket.

Figure 11 shows a profile of the interaction energy of the three template compounds, **GP1** through **GP3**, and a selected subset of designed compounds. For the primary candidates, the composition of the interaction energies associated with each of the binding regions, as well as the other residues (labeled as 'others' in the figure) that are not part of the binding site, is also displayed. On average, the binding site residues contribute over 80% of the total interaction energy. For this series of ligands, interaction with every sub-region makes a favorable contribution to binding except for a couple of cofactor components in **R6** and **R47**. Because the majority of the ligands are glucose analogs, the general characteristics of the energy profiles are similar. On the whole, the interaction energies from the two binding pockets are comparable in magnitude across the series. The contribution of interaction energy from the cofactor region is generally small, except for a few cases where there are close contacts (e.g., **R7** and **R13**). Interestingly, the gain in interaction energy due to these contacts is of-

ten offset by a less favorable contribution from the residues in the hydrophilic region. It appears the existing hydrogen bonds in the hydrophilic region are somewhat weakened by the structural adjustment of the glucose moiety that is necessary to yield better interactions with the cofactor. As a general trend, the plot shows that the increase in score can be related to a more favorable interaction energy; the correlation coefficient between the two variables is -0.62 for this series of compounds. Figure 12(a) shows a stereoview of **GP3** and three of the high ranking glucose derivatives (**R1**, **R7** and **R9**) in the binding site. It is evident that the new substituents make additional hydrogen bonds to the protein while the parent hydrogen bonding network of the glucose moiety is maintained. Figure 12(b) shows the four non-glucose candidates (**R30**, **R41**, **R47** and **R48**). In general, they form a moderate number of hydrogen bonds (between 5 and 10) to the binding site.

The interaction energy profiles for the 12 secondary compounds, shown in Figure 11, are much more variable than the primary series. It is clear that the ranking of this series of compounds no longer

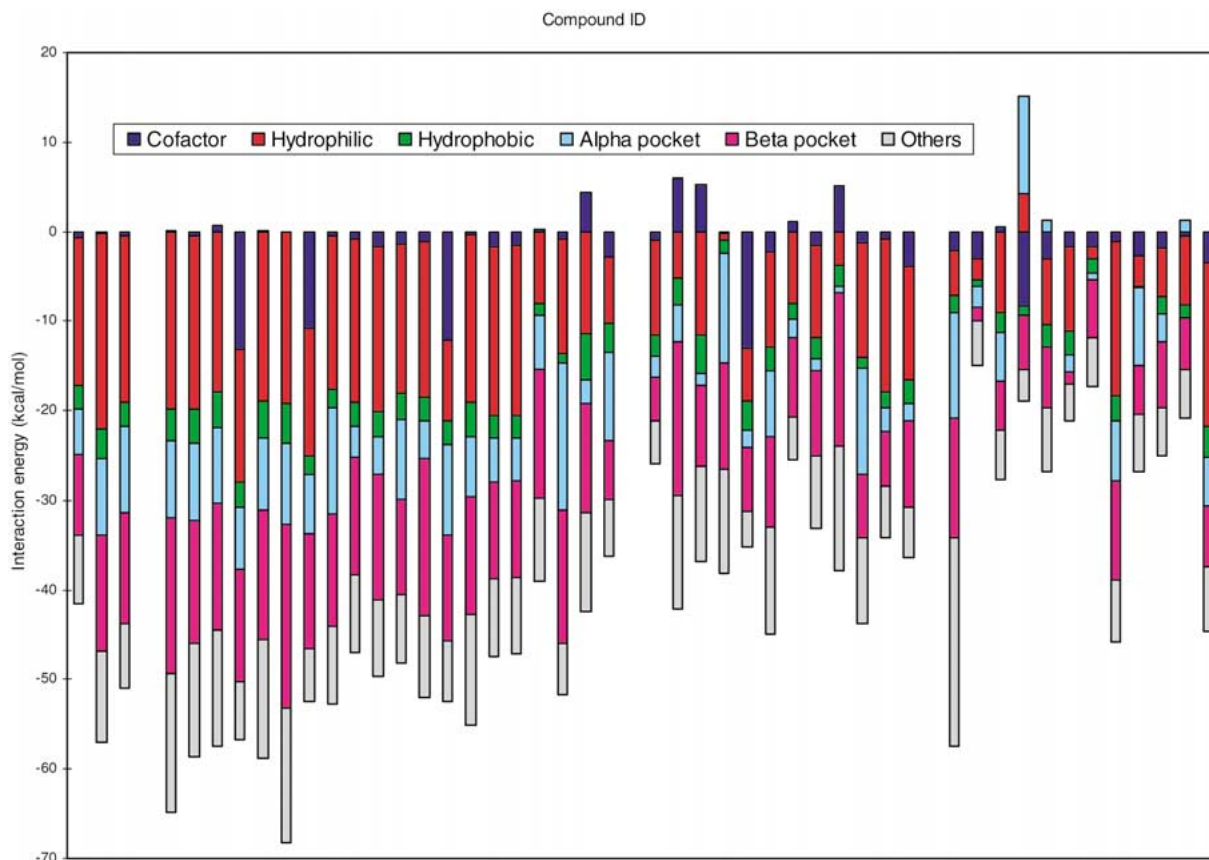


Figure 11. The interaction energy profiles for: (i) **GP1-3**; (ii) a subset of 20 primary candidates; (iii) the 12 secondary candidates; (iv) a subset of 12 low ranking compounds selected at random.

corresponds to their interaction energies; in fact, the correlation between the overall energy and the $pK_{i, \text{final}}$ is not meaningful ($r = 0.31$). Furthermore, the pattern of their energy profiles is often very different from the templates or even amongst themselves, which is a manifestation of their chemical diversity. For example, the distinctive pattern in **R53**, for which 63% of the total interaction energy is contributed by the α - and β -regions, indicates that the ligand is predicted to bind tightly to the substituent binding pocket. Figure 12(c) shows three examples of the secondary candidate positions which illustrate some of their characteristics. Generally, they make several (~ 4) hydrogen bonds with the enzyme (e.g., **R50**, **R51** and **R57**).

The remaining pool of compounds is predicted not to bind strongly to GP. The energy profiles some of these molecules are also shown in Figure 11. Like the previous series, they have a wide range of interaction energies. Some of them have favorable interaction energy. For example, **B5** is a tri-peptide that has an in-

teraction energy similar to **GP2**. However, since many of its polar groups do not pair up with polar residues in the protein, it is unlikely that the large desolvation energy could be recouped; in fact, its ΔG_{elec} value is three times less favorable than for **GP2**. Other molecules have less favorable interaction energies. Some of them do not possess the necessary functionality or the correct geometry to make satisfactory interactions. For example, **I3** and **G14** only make one or two hydrogen bonds with the protein. The poor interaction energy (-4 kcal/mol) of **I48** is due to unfavorable steric contacts with the α -region and the hydrophilic ring residues.

Concluding discussion

A total of 301 candidate ligands for GP have been designed using an array of computational approaches. Forty-nine primary and 12 secondary compounds have been established as promising candidates using an

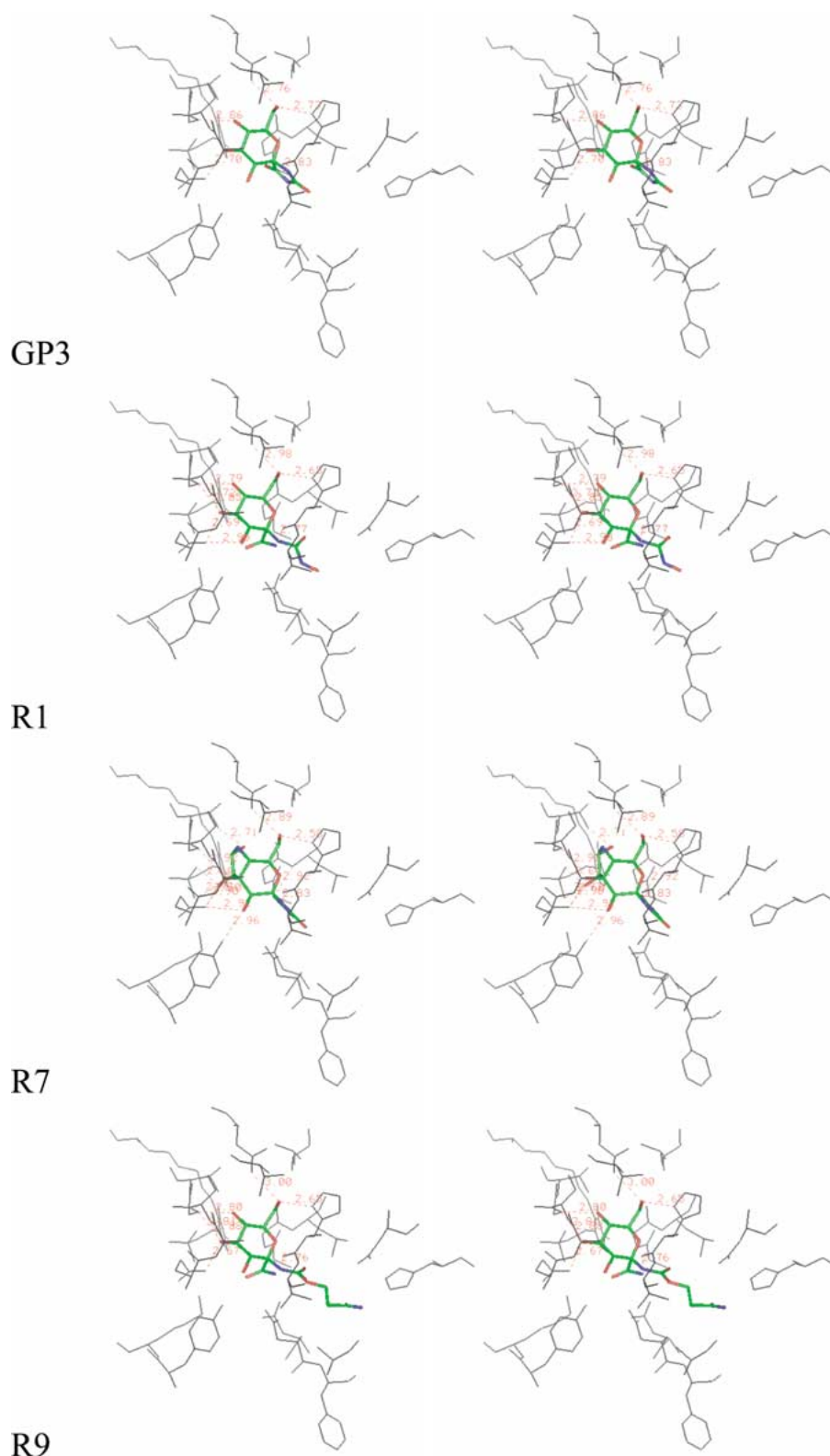


Figure 12a. Stereoview of **GP3** and the three high ranking primary candidates that are derivatives of glucose.

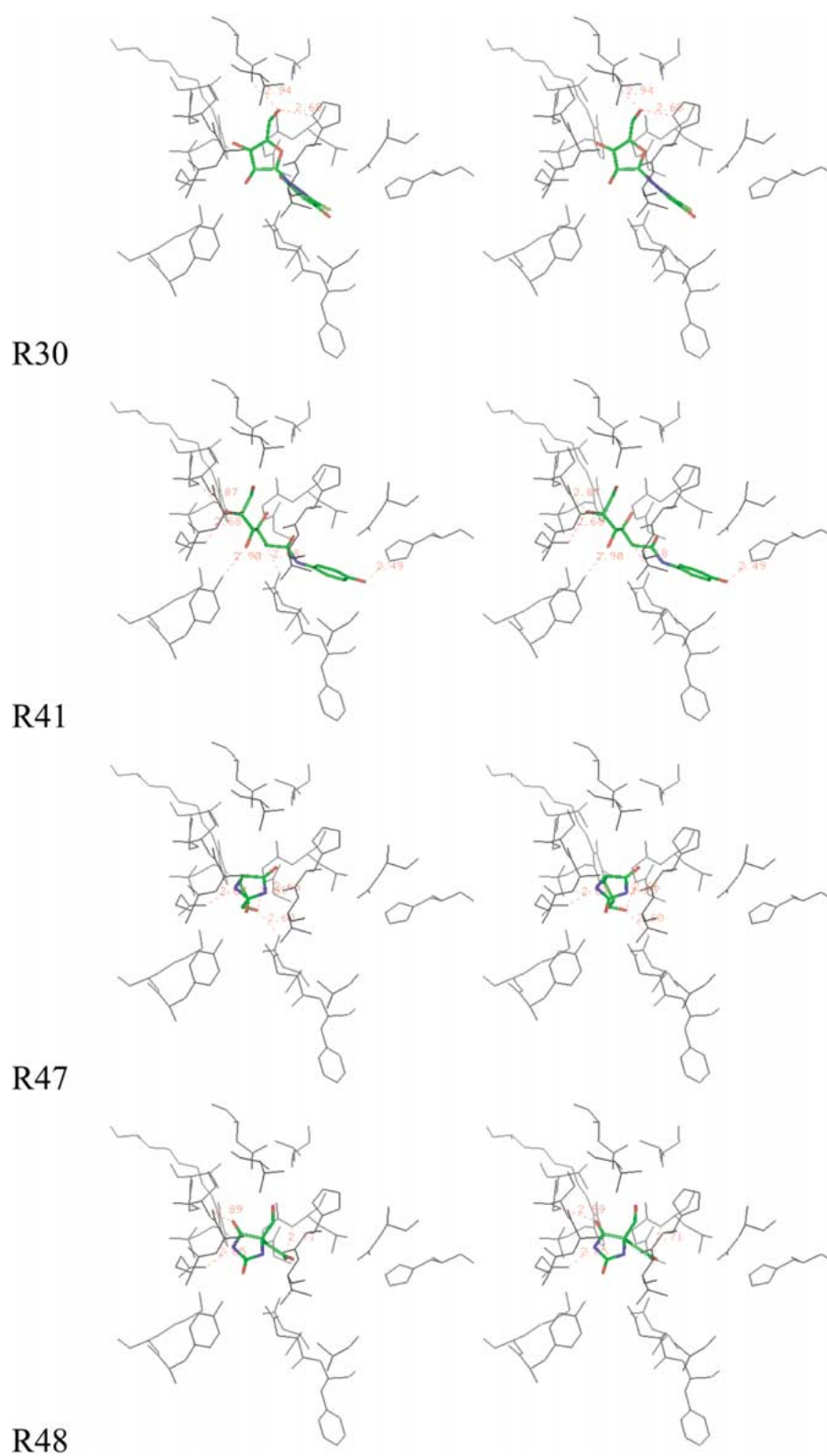


Figure 12b. Stereoview of four primary candidates that are not derivatives of glucose.

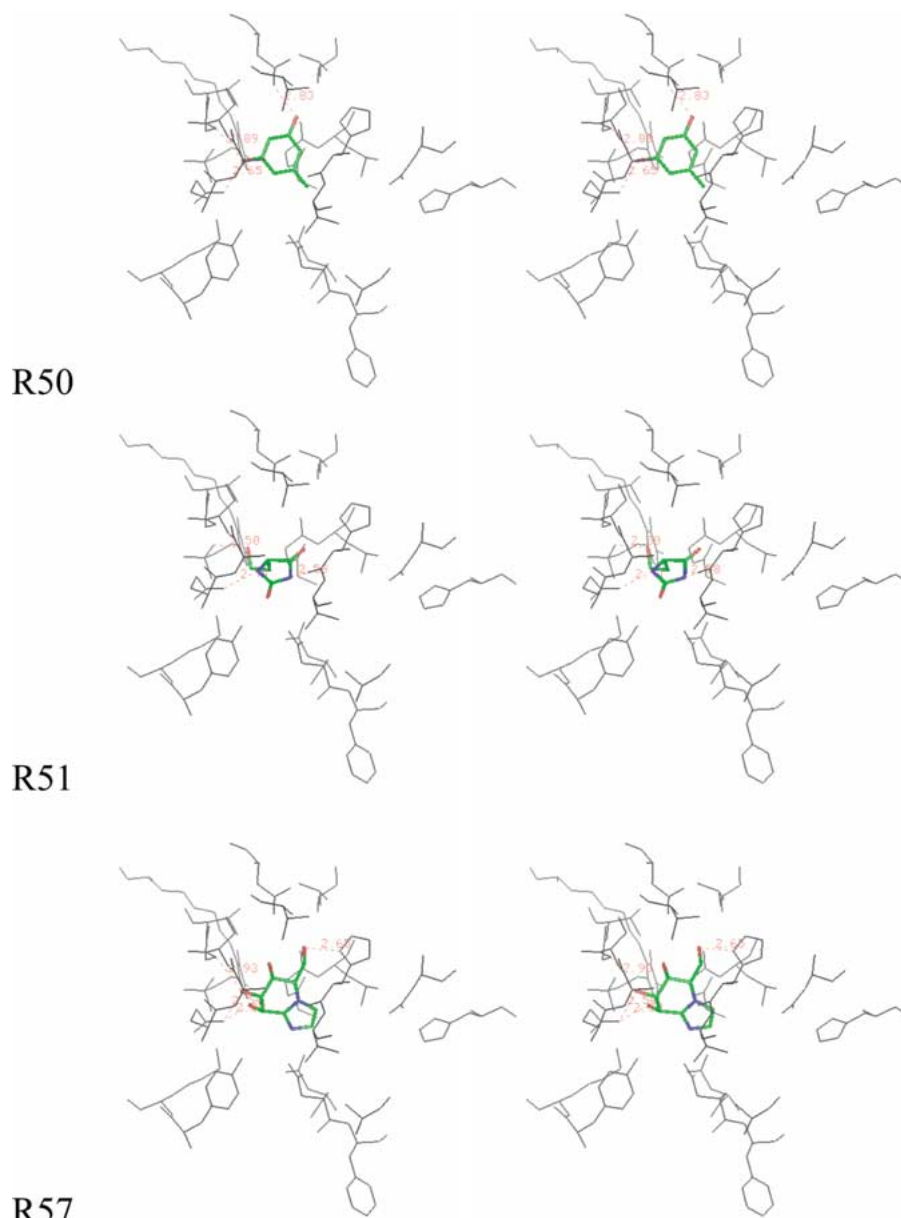


Figure 12c. Stereoview of three secondary candidates.

evaluation scheme that combines several affinity prediction models. The effectiveness of the design is demonstrated by a comparison of the prediction results with 30 randomly selected compounds, amongst which no hits have been identified.

Locally enhanced sampling (LES) methods [51], such as MCSS, provide an effective way to obtain functionality maps describing preferred binding sites and directionality of bound chemical groups for the protein. This calculation is in general straight forward,

though it becomes computationally expensive with the exploration of many different functional groups. In contrast, the analysis of such maps poses a more challenging problem because there are various ways of utilizing them. In the initial application, nine novel ligands were designed manually by connecting the functional groups to several pre-defined skeletons (six of them became primary candidates). A comprehensive visual analysis of such complexity is time consuming since a large number of potential substituent

attachments must be considered. Automated ligand design programs have been developed to facilitate this process. They include programs such as CON-NECT [15, 52], HOOK [17], dynamic ligand design (DLD) [18], computational combinatorial ligand design (CCLD) [52, 53] and OLIGO [28]; the latter was used to construct peptidomimetic compounds in this work. As evident from a growing list of ligand design applications [16, 54–58], we believe that the complementary combination of a LES method and an automated ligand connector will be of increasing importance as part of the tool kit for *de novo* drug design in future.

The LUDI program offers an alternative approach for positioning simple fragments in the binding site. To orient the molecular fragments, it relies on a pre-defined set of empirical rules which greatly improves the time performance of the program at the expense of accuracy (i.e. optimal positioning of fragments) [17, 43]. We have observed in a number of cases that the subsequent structure refinement by energy minimization introduces a major change in the ligand position, which results in a significant lowering of its LUDI score with the protein. A number of ligands (14 out of 54) that were obtained by the program did not pass the preliminary LUDI filter after energy minimization, i.e. the score had fallen below the cutoff value of 300. Despite this limitation, we believe that LUDI performed well in the ligand selection from the ACD and in the fragment attachments from its standard library. LUDI led to the suggestion of four top tier compounds.

The best results were obtained using the CoMFA/LINK approach. In fact, 32 out of the 49 primary ligands were derived with this method. However, it should be noted that they all started with three known ligands (**GPI-3**). Its computational efficiency is quite impressive: the bottleneck of the process is actually the MOPAC calculation for the AM1 partial atomic charges rather than the prediction of activity using CoMFA. For the present study, the use of the computationally demanding MOPAC charges was necessary because the standard charging methods (Gasteiger, Hückel, and Del Re) in SYBYL did not yield satisfactory QSAR. However, the result of this study shows that it is feasible to screen a very large library of compounds based on CoMFA models, particularly if less expensive methods can be used to determine the charges.

The three LINK approaches (MCSS/LUDI/CoMFA) considered in this study can be compared. At an elementary level all three methods create new inhibitors

by connecting fragments to a molecular template, though there are subtle differences concerning: (a) the type of fragments that are available; (b) the positioning of fragments; and (c) the manner in which the attachment is made. In MCSS, the number of functional groups (65 in version 2.1) in the program is limited but the users can easily add additional groups related, for example, to synthons for combinatorial chemistry library. The positions and orientations of the MCSS groups, which are conformationally flexible, are determined by the binding site based on molecular mechanics force field calculations. The visual analysis stage is time consuming due to its interactive nature, but it plays an essential role in obtaining meaningful ligands. Furthermore, since the positions of the template molecules did not enter into the functional map calculations, we find that most of the generated fragments could not be attached to the templates due to geometric constraints or unrealistic internal structure. It is of course possible to do MCSS with the template in the binding site and use a method like DLD [18] to make the connections to it. The unrealistic geometry of the newly formed ligands can be regularized by the subsequent energy minimization, though it is often at the expense of shifting the linked fragments from their original optimal positions. The use of additional linker groups, as in DLD or HOOK, may be a solution to this problem, though it adds complexity to the overall calculation. In LUDI, a total of 1163 fragment structures are available from its standard library, which can be expanded by its auxiliary utility program, GENFRA [45]. The positioning of these fragments are determined by the goodness of fit to the interaction sites of LUDI obtained by a set of geometric rules extracted from statistical behavior of the interactions of small molecules. During the search, it allows up to two fragment torsion angles to vary simultaneously at one time, using large angle increments (either 120° or 180°, depending on the hybridization of the atoms involved). The use of geometric operations and grid-based conformational search makes the computation very fast, although the fragment positions are generally less reliable than those derived from force field based calculations. Like most structure-based methods, LUDI treats the protein as rigid and it does not take induced fit into account in optimizing the interactions. In the LINK model of LUDI, the fragments are positioned where they can be connected to the templates at the pre-defined attachment points. Subsequently, the ligand building process can be performed in a semi-automatic manner. A feature,

Table 5. The calculated scores (according to Equation 2), predicted pK_i (Equation 3) and experimental pK_i values for seven H compounds. **H7**, **H15** and **H32** were identified as the secondary candidates from virtual screening calculations prior to experimental measurements.

Compound	Score	Predicted pK_i	Experimental pK_i
H4	-2.89	1.58	2.15
H7	-0.99	2.53	<1.82
H14	-2.28	1.88	no inhibition @ 10 mM
H15	-1.63	2.21	2.70
H20	-2.27	1.89	2.35
H32	-1.46	2.29	no inhibition @ 10 mM
H34	-2.45	1.80	no inhibition @ 10 mM

not observed in this study, is that LUDI permits multiple linkage of the same fragment to the template to form additional ring system. Tripos Inc. supplies three molecular spreadsheets (sub1v.tbl, sub2v.tbl and sub3v.tbl), which contain a total of 383 fragment structures (309 single-, 60 double- and 14 triple-linkage fragments), as the standard substituent library for the CoMFA application. As for the other methods, the list of fragments can be readily customized and expanded. Using a standard conformation, a new fragment is attached in a trans configuration to the pre-defined linkage points, and the predicted binding affinity of the new ligands is evaluated by an empirically derived QSAR model. A disadvantage of this approach is that it is analog-based. Since the initial positions and conformation of the new fragments are obtained by energy minimization without the protein, it is possible that the generated coordinates of the new analog may be unrealistic (i.e., it may have severe steric clashes with the binding site when it is placed into the protein). An improved protocol would involve incorporation of binding site information to derive a more realistic conformation of the ligand at the prediction stage. In summary, all three approaches are useful methods that can introduce novel functional groups for existing lead molecules, and depending on the problem, one or another method will be more successful.

We have considered two screening approaches to identify promising leads from the NCI library. The first is a 2D method that makes use of a previously derived HQSAR model. The search process is remarkably fast owing to the uncomplicated, yet informative, fingerprint system used for the query. A major shortcoming of the method is that for subsequent evaluations, such as the detailed activity pre-

diction used here, substantial effort would be required to dock the hits into the 3D binding site. The second method makes use of 3D molecular similarity, by invoking a SMGNN QSAR model to rank the database compounds in a quantitative fashion. Because this system makes explicit use of 3D information, the computation is more demanding than the previous approach, although in our experience this is usually compensated for by a simpler post processing stage; i.e., introducing a SMGNN ligand into the binding site is generally easier than a corresponding HQSAR candidate. We identified four potential candidates from the HQSAR screen and a single candidate from the SMGNN screen. It should be noted that in both cases the queries were derived from molecular similarity and the hits were ranked based on QSAR predictions. Thus, it would be interesting to complement the present effort using a pharmacophore-based query. Commercial programs such as CATALYST and UNITY, which perform 3D database searches, can be used for this kind of application.

None of the 36 SDB compounds were recognized as a primary lead molecule; the best scores are for **H7**, **H32** and **H15**, which are included as secondary candidates. Two factors may have contributed to their relatively low predicted pK_i values. As already discussed, the activity prediction system contains a degree of bias against novel chemical classes, other than glucose-like molecules. Furthermore, docking these molecules into the binding site was, in itself, a challenging task. The docking procedure used in this study explored only very limited possibilities. It is possible that some of docked conformations are sub-optimal and may result in a lowering of predicted activity. The inhibition activities of seven compounds in this set have now been determined (K.M. Watson, private communication). The experimental pK_i values are listed in Table 5 together with the predicted activities (as determined by Equations 2 and 3) from virtual screening calculations. There appears to be no correlation. This is not surprising, given that the rms error of prediction is approximately 0.5 log unit (for the 30 training compounds) and the range of predicted pK_i values (from 1.58 to 2.53) for the seven compounds span less than a log unit. On an absolute scale, however, the predictions seem quite accurate; none of the compounds that were measured have potency better than 1 mM (i.e., $pK_i = 3$), which is in accord with the predictions. Furthermore, the most potent compound (**H15**) in this set is in fact a secondary candidate identified by the virtual screening process. It has a measured pK_i value

Table 6. Comparison of the different methods in this study. Method (a) MCSS/LINK; (b) MCSS/OLIGO; (c) LUDI/ACD; (d) LUDI/LINK; (e) CoMFA/LINK; (f) HQSAR/NCI; (g) SMGNN/NCI; (h) Docking/SDB; (i) Docking/CombiChem.

	Method								
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
Method class ^a	S	S	D/S	D/S	L	D/L	D/L	D/S	S
Computational cost in generation ^b	3	3	3	1	2	1	2	1	1
Human intervention ^c	3	2	1	2	2	1	1	1	1
Post processing ^d	1	1	1	1	2	3	2	3	2
Compound availability ^e	3	1	1	3	3	2	2	1	2
Top candidate no.	A8	B11	C4	D41	E25	F4	G20	H7	I16
Rank ^f	7	185	50	9	1	18	30	62	181
2nd best candidate no.	A7	B14	C1	D44	E28	F2	G19	H32	I3
Rank	14	196	53	13	2	24	57	70	190
3rd best candidate no.	A9	B5	C3	D4	E17	F8	G18	H15	I5
Rank	15	198	73	23	3	39	66	71	195

^aD = database search; L = ligand-based; S = structure-based.

^b1 = fast (<1 CPU h); 2 = moderate; 3 = slow (> 12 CPU h).

^c1 = little; 2 = moderate; 3 = intensive.

^d1 = little; 2 = moderate; 3 = intensive.

^e1 = readily available from vendor; 2 = easy synthesis; 3 = more challenging synthesis.

^fRanking out of 301 candidates.

of 2.70, which is almost as potent as α -glucose (2.77), the physiological regulator of GP. From a drug design perspective this is very encouraging because **H15** is a low molecular weight non-glucose compound that could serve as a lead for future optimization.

For reasons similar to the above, none of the 75 CombiChem compounds were predicted to bind favorably with GP. This is not surprising since a simple design approach was used to construct this combinatorial library and many of compounds are very similar. This application resembles, in reality, a random screening effort. For the purpose of lead identification, what is needed is a heuristic way to design a library whose members are maximally diverse, as determined by some chemical or physical criteria, and, ideally made up of substituents that are chosen for the specific binding site by MCSS or related methods. Additional constraints could be imposed to achieve such diversity. They could include, for example, the in-house availability of the agents, the cost of the starting materials, the drug-likeness of the products [59, 60], or even the ease of experimental identification of the lead molecule [61]. A DirectedDiversity approach [62], which integrates computational design with combinatorial chemistry and high throughput screening, is likely to be an important future direction in drug design.

A schematic comparison of the ligand design methods considered in this paper is made in Table 6. The first entry is a classification of computational approach, i.e., database search, ligand- or structure-based method, or a combination of two. The next four attributes are related to their performance, including (i) the computational cost in ligand generation; (ii) the amount of human intervention required to select or construct candidate ligands; (iii) the degree of post processing needed (e.g., extensive structure refinement or manual ligand docking); (iv) the availability of the candidate ligands; (v) the overall ranking of their top three candidates. It is clear that each method has its own merits and shortcomings, and no single approach has outperformed other methods in every category. Also, it is evident that the structural diversity of the new inhibitors is bounded by the underlying algorithm and the evaluation strategy inherent to each design method. For example, the CoMFA/LINK approach tends to suggest conservative modifications of lead structures whereas the LUDI/ACD method yields novel templates. Thus, the exploration of many different computational ligand design strategies will potentially lead to a richer collection of lead structures.

Some limitations of the present ligand design strategy are discussed here. First, the flexibility of the protein has not been taken into account in this study.

The structure-based programs, MCSS and LUDI, construct new inhibitors based on a rigid protein structure. It is possible to run multiple calculations using many protein conformations, which may be derived computationally from a dynamics simulation, or experimentally from different ligand-protein complex structures. Alternatively, after a candidate ligand has been positioned in the binding site, one may allow the protein to relax together with the ligand during structural refinement. Second, although the computational tools used in this study are already quite sophisticated, it is clear that human inference still plays an important role in the design. For example, the evaluation of chemical stability and synthetic tractability of the new entities are very challenging computational problems and are best handled by a human expert. To alleviate this task, it will be useful to establish a rule-based filter to flag compounds that contain substructures known to be reactive or problematic, such as Michael acceptors or anilines. Third, the utility of many different software packages for the design and evaluation of new ligands would require efficient data handling. Ideally, the data transfer between the programs should be fully automated so that this time-consuming interactive tasks can be eliminated. Last but not least, one must recognize the huge gulf between a tightly bound inhibitor and a drug [63]. Far too often, promising candidates are abandoned during clinical trials for a variety of reasons, including low bioavailability, high toxicity or poor pharmacokinetics. Thus, these factors should be introduced earlier during lead identification and optimization [64]. In this regard, scoring functions that incorporate *in silico* prediction for absorption [65, 66], metabolism [67–69], bioavailability [70], and toxicity [71] will have significant impact in future drug design.

In conclusion, we emphasize that drug design is an iterative process. We are hopeful that, with rapid advances in experimental technologies and increasingly sophisticated computational tools, a potent bioavailable drug that inhibits GP will be designed to treat diabetes in the not-too-distant future.

Acknowledgements

We thank Molecular Simulations Inc. and Tripos Inc. for the software support. This work is supported in part by a GOELI grant from the National Science Foundation and a gift from Eli Lilly and Company. We are grateful to Dr. Kim Watson and Prof. Louise Johnson for providing the coordinates of the GP x-

ray structures and the data on the SDB compounds, after completion of this work. We thank Prof. David Austin for help discussions on combinatorial library synthesis.

References

1. Watson, K.A., Mitchell, E.P., Johnson, L.N., Son, J.C., Bichard, C.J., Orchard, M.G., Fleet, G.W., Oikonomakos, N.G., Leonidas, D.D. and Kontou, M., *Biochemistry*, 33 (1994) 5745.
2. Watson, K.A., Mitchell, E.P., Johnson, L.N., Cruciani, G., Son, J.C., Bichard, C.J.F., Fleet, G.W.J., Oikonomakos, N.G., Kontou, M. and Zographos, S.E., *Acta Cryst.*, D51 (1995) 458.
3. Kuntz, I.D., *Science*, 257 (1992) 1078.
4. van Gunsteren, W.F., King, P.M. and Mark, A.E., *Q. Rev. of Biophys.*, 27 (1994) 435.
5. Good, A.C. and Mason, J.S., In Lipkowitz, K.B. and Boyd, D.B. (eds), *Reviews in Computational Chemistry*, 1996, VCH Publishers, New York, NY, p. 67.
6. Grigorov, M., Weber, J., Tronchet, J.M., Jefford, C.W., Mihous, W.K. and Maric, D., *J. Chem. Inf. Comput. Sci.*, 37 (1997) 124.
7. Kaminski, J.J., Rane, D.F., Snow, M.E., Weber, L., Rothofsky, M.L., Anderson, S.D. and Lin, S.L., *J. Med. Chem.*, 40 (1997) 4103.
8. Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B. and Marshall, G.R., *J. Comput. Aid. Mol. Des.*, 3 (1989) 3.
9. Langlois, M., Bremont, B., Rousselle, D. and Gaudy, F., *Eur. J. Pharmacol.*, 244 (1993) 77.
10. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., *Science*, 259 (1993) 1445.
11. Gschwend, D.A., Sirawaraporn, W., Santi, D.V. and Kuntz, I.D., *Prot. Struct. Funct. Genet.*, 29 (1997) 59.
12. Böhm, H.-J., *J. Comput. Aid. Mol. Des.*, 6 (1992) 593.
13. Böhm, H.-J., *J. Comput. Aid. Mol. Des.*, 8 (1994) 243.
14. Miranker, A. and Karplus, M., *Prot. Struct. Funct. Genet.*, 11 (1991) 29.
15. Caffisch, A., Miranker, A. and Karplus, M., *J. Med. Chem.*, 36 (1993) 2142.
16. Joseph-McCarthy, D., Hogle, J.M. and Karplus, M., *Prot. Struct. Funct. Genet.*, 29 (1997) 32.
17. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., *Prot. Struct. Funct. Genet.*, 19 (1994) 199.
18. Miranker, A. and Karplus, M., *Prot. Struct. Funct. Genet.*, 23 (1995) 472.
19. Rotstein, S.H. and Murcko, M.A., *J. Comput. Aid. Mol. Des.*, 7 (1993) 23.
20. Gillet, V., Johnson, A.P., Mata, P., Sike, S. and Williams, P., *J. Comput. Aid. Mol. Des.*, 7 (1993) 127.
21. Gillet, V.J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z. and Johnson, A.P., *J. Chem. Inf. Comput. Sci.*, 34 (1994) 207.
22. Gerber, P.R. and Müller, K., *J. Comput. Aid. Mol. Des.*, 9 (1995) 251.
23. Moon, J.B. and Howe, W.J., *Prot. Struct. Funct. Genet.*, 11 (1991) 314.
24. DeWitte, R.S. and Shakhnovich, E.I., *J. Am. Chem. Soc.*, 118 (1996) 11733.
25. DeWitte, R.S., Ishchenko, A.V. and Shakhnovich, E.I., *J. Am. Chem. Soc.*, 119 (1997) 4608.

26. MCSS Version 2.1, Evensen, E.R., Joseph-McCarthy, D. and Karplus, M. Harvard University, Cambridge, MA.
27. Cramer, R.D., III, Patterson, D.E. and Bunce, J.D., *J. Am. Chem. Soc.*, 110 (1988) 5959.
28. OLIGO Version 0.1, developed by Evenson, E.R. and Karplus M. Harvard University, Cambridge, MA.
29. Muegge, I. and Martin, Y.C., *J. Med. Chem.*, 42 (1999) 791.
30. Åqvist, J., Medina, C. and Samuelsson, J.E., *Prot. Eng.*, 7 (1994) 385.
31. Charifson, P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P., *J. Med. Chem.*, 42 (1999) 5100.
32. So, S.-S. and Karplus, M., *J. Comput. Aid. Mol. Des.*, 13 (1999) 243.
33. HQSAR Version 1.0 Tripos, Inc., St. Louis, MO.
34. So, S.-S. and Karplus, M., *J. Med. Chem.*, 40 (1997) 4347.
35. So, S.-S. and Karplus, M., *J. Med. Chem.*, 40 (1997) 4360.
36. So, S.-S. and Karplus, M., *J. Med. Chem.*, 39 (1996) 1521.
37. So, S.-S. and Karplus, M., *J. Med. Chem.*, 39 (1996) 5246.
38. Hahn, M., *J. Med. Chem.*, 38 (1995) 2080.
39. Hahn, M. and Rogers, D., *J. Med. Chem.*, 38 (1995) 2091.
40. Andrea, T.A. and Kalayeh, H., *J. Med. Chem.*, 34 (1991) 2824.
41. So, S.-S. and Richards, W.G., *J. Med. Chem.*, 35 (1992) 3201.
42. Neria, E., Fischer, S. and Karplus, M., *J. Chem. Phys.*, 105 (1996) 1902.
43. Caflisch, A. and Karplus, M., *Perspect. Drug Discov. Des.*, 3 (1995) 51.
44. Böhm, H.-J., *J. Comput. Aid. Mol. Des.*, 6 (1992) 61.
45. Böhm, H.-J., *J. Comput. Aid. Mol. Des.*, 8 (1994) 623.
46. Fisanick, W., Lipkus, A.H. and Rusinko, A.I., *J. Chem. Inf. Comput. Sci.*, 1994 (1994) 130.
47. Sadowski, J. and Gasteiger, J., *Chem. Rev.*, 7 (1993) 2567.
48. INSIGHT Version 95.0 Molecular Simulations Inc., San Diego, CA.
49. Cerius² Version 3.0 Molecular Simulations Inc, San Diego, CA.
50. Schaefer, M. and Karplus, M., *J. Phys. Chem.*, 100 (1996) 1578.
51. Elber, R. and Karplus, M., *J. Am. Chem. Soc.*, 112 (1990) 9161.
52. Caflisch, A., *J. Comput. Aid. Mol. Des.*, 10 (1996) 372.
53. Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C. and Caflisch, A., *Prot. Struct. Funct. Genet.*, 37 (1999) 88.
54. Caflisch, A. and Karplus, M., *Proc. Natl. Acad. Sci. U.S.A.*, 91 (1994) 1746.
55. Grootenhuis, P.D.J. and Karplus, M., *J. Comput. Aid. Mol. Des.*, 10 (1996) 1.
56. Leclerc, F. and Karplus, M., *Theor. Chem. Acc.*, 101 (1999) 131.
57. Caflisch, A., Schramm, H.J. and Karplus, M., *J. Comput. Aid. Mol. Des.*, 14 (2000) 161.
58. Stultz, C.M. and Karplus, M., *Prot. Struct. Funct. Genet.*, 40 (2000) 258.
59. Ajay, Walters, P. and Murcko, M.A., *J. Med. Chem.*, 41 (1998) 3314.
60. Sadowski, J. and Kubinyi, H., *J. Med. Chem.*, (1998) .
61. Brown, R.D. and Martin, Y.C., *J. Med. Chem.*, 40 (1997) 2304.
62. Agrafiotis, D.K., *J. Chem. Info. Comput. Sci.*, 37 (1997) 841.
63. Verlinde, C.L.M.J. and Hol, W.G.J., *Structure*, 2 (1994) 577.
64. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., *Adv. Drug Delivery Rev.*, 23 (1997) 3.
65. Wessel, M.D., Jurs, P.C., Tolani, J.W. and Muskal, S.M., *J. Chem. Inf. Comput. Sci.*, 38 (1998) 726.
66. Palm, K., Kuthman, K., Ungell, A.L., Strandlund, G., Beigi, F., Lundahl, P. and Artursson, P., *J. Med. Chem.*, 41 (1998) 5382.
67. Lewis, D.F., Dickins, M., Lake, B.G., Eddershaw, P.J., Tarbit, M.H. and Goldfarb, P.S., *Toxicology*, 133 (1999) 1.
68. de Groot, M.J., Ackland, M.J., Horne, V.A., Alex, A.A. and Jones, B.C., *J. Med. Chem.*, 42 (1999) 4062.
69. de Groot, M.J., Ackland, M.J., Horne, V.A., Alex, A.A. and Jones, B.C., *J. Med. Chem.*, 42 (1999) 1515.
70. GastricPlus Version 1.3.3 Simulations Plus, Inc., Lancaster, CA.
71. Richard, A.M., *Mutat. Res.*, 400 (1998) 493.