

SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence

Robert S. DeWitte and Eugene I. Shakhnovich*

Contribution from the Department of Chemistry and Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

Received March 7, 1996[⊗]

Abstract: In this paper, we present SMoG (Small Molecule Growth), a novel, straightforward method for *de novo* lead design and the evidence for its effectiveness. It is based on a simple model for ligand-protein interactions and a scoring that is directly related to the free energy through a knowledge-based potential. A large number of structures are examined by an efficient metropolis Monte Carlo molecular growth algorithm that generates molecules through the adjoining of functional groups directly in the binding region. Thus SMoG is a method that is able to rank a large number of potential compounds according to binding free energy in a *short* time. In this sense, SMoG represents a step toward an ideal computational tool for ligand design.

Introduction

The General Problem. Structure-based drug design is beginning to play an important role in the discovery of new therapeutic molecules. Particularly when a lead compound is already known, and the three dimensional structure of the protein-ligand complex has been determined, computer modeling provides an opportunity for assessment of the feasibility of related compounds as ligands. When no lead compound is known, however, computers can help the medicinal chemist by sampling a large variation of structures quickly and thoroughly, generating potential lead candidates automatically. The *ideal* computational tool for this *de novo* lead design will be able to test *many* structures in a *short* period of time and arrange them into a ranked list based on an *accurate prediction of binding free energies*, since the latter reflect actual binding propensities. Of course, the terms *many* and *short* need to be clarified.

The number of structures that need to be sampled can be determined from an analysis of the combinatorics of molecular architecture. Let us consider, for the sake of argument that we are in pursuit of chemicals on the order of 300 atomic mass units or roughly 20–25 second period elements. If we consider small organic molecules as simple combinations of functional groups, candidate molecules can be thought of as the choice and arrangement of approximately five functional groups. Even a library of functional groups as rudimentary as the one presented in this paper (see Table 1) contains about 50 fragments, resulting in the overwhelming number of structures to be considered being 50⁵ or over 10⁸ candidates. Obviously, this figure grows very quickly as the diversity of possible structures is increased by enlarging the library of molecular fragments. Moreover, the number of arrangements and conformations of five fragments can be estimated in the tens of thousands. These arguments lead us to define the term *many* as on the order of billion or trillion. It is clear that exhaustive searches over all possible compounds are beyond the limit of practical computation, so in order to enhance practicability, searches must be tuned to specific traits or the library of chemical fragments needs to be trimmed.

Programs that perform analysis instantaneously can be used interactively and thereby present the chemist with novel insight

Table 1. Fragments Used in the Small Molecule Growth Algorithm

amide	cyclohexene	methyl	pyrimidine
amine	1,2-dithian	<i>n</i> -butyl ^a	pyridine
carbonyl	ethane	naphthalene	pyrrole
carboxylic acid	ethene	nitrile	sulfate
chloride	fluoride	nitro	sulfide
cyanide	furan	phenyl	<i>tert</i> -butyl
cyclooctane ^a	glucose ^a	phosphate	tetrahydrofuranlyl
cyclopentane	hydroxyl	propane	tetrahydrothienyl
cycloheptane ^a	indole	propene	thiophene
cyclohexane ^a	iodide	purine	trifluoromethyl

^a Indicates that multiple conformations are represented.

in real time. This is not necessary, however, as most researchers are comfortable with computer programs that require a day's computation. However, computations that extend beyond a day become less desirable tools.

As we will point out in the discussion section, the several computational tools that have been presented in the literature to date have made excellent use of algorithmic programming in order to overcome, in part, the combinatorial problem discussed above. However it is with respect to the *prediction of free energies of binding* that today's methods are most weak. The most widely accepted approach for computational estimation of free energy of binding involves sophisticated simulations using an empirical potential and stepwise estimation of the changes in enthalpy and entropy at each stage in a thermodynamic cycle. These calculations demand on the order of days of computation for each ligand candidate. As a result, the present methods rely on alternative scoring techniques to provide a short list of candidates for complete thermodynamic determination or, involving even greater expense, chemical synthesis and experimental determination of binding properties. Each of these scoring methods in some way approximates the binding free energy, whose explicit form is unknown, but which must contain the following terms (*g* being free energy, *e* being energy, and *s* being entropy):

$$\begin{aligned} \Delta g_{\text{binding}} &= \Delta e_{\text{binding}} - T\Delta s_{\text{binding}} \quad (1) \\ &= \Delta e_{\text{complex formation}} - T\Delta s_{\text{complex formation}} + \\ &\quad \Delta e_{\text{solvation/desolvation}} - T\Delta s_{\text{solvation/desolvation}} \end{aligned}$$

where the terms marked complex formation refer to the associ-

* Address correspondence to E. I. Shakhnovich, Dept. of Chemistry, Harvard University, 12 Oxford Street Cambridge, MA 02138. Phone: (617)495-4130. Fax: (617)496-5948. Email: eugene@diamond.harvard.edu.

[⊗] Abstract published in *Advance ACS Abstracts*, November 1, 1996.

ation event between ligand and receptor *in vacuo*, and those marked solvation/desolvation refer to the effects explicitly due to solvation. Hence $\Delta e_{\text{binding}}$ refers to the interaction energy minus any intramolecular strain induced upon complex formation, and $\Delta s_{\text{binding}}$ refers to the change in conformational freedom induced by formation of the complex. Energies of solvation are those energetic factors arising from the transfer of hydrophilic and lipophilic groups from aqueous solvent to the more lipophilic region of the protein binding site, and entropy of solvation refers to changes in the order of the solvent at the interface between ligand and solvent and protein and solvent upon complex formation. In many of the quantitative approximations to the full expression for binding free energy that have been implemented in the past for the purpose of ligand design, the scoring is based solely on the interaction energy between the ligand and protein in the complex as the single most important contributor to the free energy. In other schemes based more on spatial complementarity than chemical complementarity, the *ansatz* that solvation contributions are proportional to exposed surface area motivate the scoring strategies. In both of these approximations, the scoring system is rather incomplete, unfortunately resulting in erroneous ranking of the candidate ligands. Hence, great progress in computational ligand design can be achieved with the introduction of an improved evaluation of binding free energy which is as efficient as the approximations currently in use. A recent paper¹ has presented a significant step in this direction through the application of a knowledge-based potential to an interaction model based on shared surface area. By adjusting two free parameters in their model, predicted free energies can be fit to the experimental binding free energies of a set of known, related ligands to reasonable accuracy. By the careful choice of a contact-based interaction model, our interaction potential reflects the trends in binding free energy without free parameters, thus eliminating the need for a series of known related ligands in the hunt for a lead compound.

Coarse-Graining and the Knowledge-Based Potential. In order to overcome this limitation and therefore provide a more directly predictive *de novo* design tool, we implement here a coarse-grained model with a corresponding knowledge-based potential. Whereas the details of our implementation will be described in the methods section, it is relevant to introduce the nature of our approximations at this stage in order to shed light on how this novel method provides an approximate description of the binding free energy, incorporating effects from all of the terms of eq 1. The model we employ is intermediate between crude functional group or amino acid representations of chemical structure and traditional molecular dynamics force fields. Our model treats both ligand and protein in an all-atom representation but assumes a simplified form of their interaction.

According to arguments made by Finkelstein and co-workers² one can apply the principles of canonical statistical mechanics to subsets of proteins in that tiny subsets of a folded protein are in thermal equilibrium with each other. This implies that the information present in crystal structures of proteins and crystal structures of protein-ligand complexes (insofar as the lifetime of the complexed form is significantly longer than the time scales of the thermal fluctuations of the system) can be disassembled into constituent pieces, and the contribution of each piece can be assigned on the basis of probabilities. This is the heart of the knowledge-based approach^{1,3} (*i.e.*: learning the interaction energies by training on a database). In this

application the database contains crystal structures of protein-ligand complexes as described below.

More formally, the postulate of equal *a priori* probabilities states that any two states at the same energy have equal probability of occupation, hence

$$p_{ij}^e = \frac{\exp\left[\frac{-e_{ij}}{kT}\right]}{Z} \quad (2)$$

where here we denote the subscripts *i* and *j* to mean different atoms on the protein and ligand, respectively, so that e_{ij} and p_{ij}^e refer to the energy and probability of an interaction between protein atom *i* and ligand atom *j*. The situation in the formation of protein ligand complexes, however, is that not all configurations of the same energy are equally likely, because of two entropic effects that arise from the strong presence of a boundary in the space sampled by the ligand. These are solvent ordering (at the protein-solvent interface, ligand-solvent interface and complex-solvent interface), and the restrictions on atomic interactions are due to steric hindrance and the nearly fixed chemical structures of the ligand and protein (*i.e.*: fixed molecular architecture and small amount of conformational freedom). These entropic effects are not correlated to the energetic events, and so we can express the total probability as a product of p^e above and a sampling probability p^s , which we can relate to a notion of entropy as

$$p_{ij}^s = \frac{\exp\left[-\frac{s_{ij}}{k}\right]}{Z} \quad (3)$$

giving a relation between probability and a notion of free energy that is dependent on the model chosen to describe the atomic interactions:

$$p_{ij} = p_{ij}^e p_{ij}^s = \frac{\exp\left[-\frac{e_{ij} - Ts_{ij}}{kT}\right]}{Z} = \frac{\exp\left[-\frac{g_{ij}^*}{kT}\right]}{Z} \quad (4)$$

which can be inverted to give an expression for g_{ij}^* from the frequency of observed interactions.

$$g_{ij}^* = -kT \log(p_{ij}) - \log(Z) \quad (5)$$

By an appropriate choice of a reference state, the partition function can be eliminated

$$\bar{g} = -kT \log(\bar{p}) - \log(Z) \quad (6)$$

$$g_{ij} = g_{ij}^* - \bar{g} \quad (7)$$

$$g_{ij} = -kT \log\left[\frac{p_{ij}}{\bar{p}}\right] \quad (8)$$

which gives a method to relate the statistical information about interatomic interactions in crystal structures of protein-ligand complexes to a two-body parameter that is a notion of free energy. By an appropriate choice of model for atomic interactions and definition of reference state, it is possible to construct the parameters g_{ij} so that their sum is an approximation of the complete form of the free energy in eq 1.

Without entering into the details here, it should be made clear that the choice of the interaction model is intrinsically a choice of length scales. We must determine the reasonable distances

(1) Wallqvist, A.; Jernigan, R. L.; Covell, D. G. *Protein Science* **1995**, *4*, 1881–1903.

(2) Finkelstein, A. V.; Gutin, A. M.; Badretinov, A. Y. *FEBS* **1993**, *325*, 23–28.

(3) Miyazawa, S.; Jerniga, R. L. *Macromolecules* **1985**, *18*, 534–552.

over which atoms project their chemical properties in order to accumulate the relevant statistics and apply the right model. Hence, knowledge-based potentials lend themselves naturally to coarse-graining techniques, where potential energy surfaces are smoothed by averaging all phenomena occurring below a cutoff length scale into properties describing the system at the specified length.

Coarse-Graining and the Search Algorithm. In principle, the combinatorial search space for molecular growth or docking algorithms is a rough energy landscape. Searching such a landscape requires careful algorithms and long search times. Fortunately, however, the identification of candidate lead molecules is not a search for the *lowest* free energy complex but rather a *low* free energy complex (or several). Still, the search is a difficult process because of the multiple minimum problem. If the search space can be made more smooth by coarse graining, however, the searching method need not be as sophisticated. For this reason, SMoG employs a metropolis Monte Carlo growth algorithm. Such a search procedure quickly samples the configuration space and the molecular space under the bias of the interaction potential (knowledge-based energy in this case). In a coarse-grained ligand design search space, a simple, hasty, search algorithm such as the one presented here can do very well in finding low energy configurations.

Measures of Success of Lead Design Methods. It is difficult to define the success of a de novo design effort in the absence of an example of a ligand that was synthesized and tested solely on the grounds of a computational tool. Only then do we have *proof of concept*. *Feasibility of concept*, can be established by several means, however, and most methods in the literature to date have been able to present several qualitatively interesting suggestions for novel ligands or the improvement of known ligands. It is unclear, however, from their conclusions whether these new candidates actually have lower binding free energies (either experimental or theoretically calculated). Therefore, in this work, we have chosen to demonstrate the ability of SMoG to predict the relative binding free energies of a series of known ligands. It is this success that gives us confidence that SMoG's combination of coarse-graining, knowledge-based potential and Monte Carlo growth algorithm provides an exciting new contribution to the search for novel pharmaceutical leads. As an example of the rich molecules that SMoG is able to produce, we do include a discussion of one design effort: a binding pocket on the CD4 protein. Greater development of the general SMoG design methodology will be left to the second paper in this series, which is forthcoming. Furthermore, we are also presently pursuing genuine *proof of concept* in collaboration with medicinal chemists.

Methods

Model and Interaction Potential. The correct model and reference state for the application of a knowledge-based potential to the protein-ligand binding event can be deduced with respect to the physical origins of the various terms in eq 1.

Changes in solvation entropy upon complex formation arise due to the loss or gain of solvent order. This solvent order is manifest as a correlation in the potential surface of the solvent exposed atoms in the ligand, protein, or complex. These correlations extend on the order of twice the size of a water molecule beyond the boundary of the ligand, protein, or complex. As the interactions formed between ligand and protein upon complexation have resulted in desolvation, there has been a change in the configurational entropy. In other words, in order to form a particular intermolecular contact, each of the atoms in contact must have been desolvated. Therefore, where much order has been destroyed, there is an entropic increase due to

desolvation for formation of that particular contact. By choosing the interaction radius between protein and ligand to be the correlation length of solvent ordering, the probabilities of the specific contacts observed will include the effect of an average over the contribution of solvation entropy to the free energy. For this reason, a simple radius of 5 Å has been chosen for our interaction model: a ligand atom is in contact with a protein atom if they lie within 5 Å of each other.

The formation of each contact also involves energetic costs for desolvation. This effect can be taken into account by the reference state. Choose

$$\bar{p} = \frac{1}{N} \sum_{ij} p_{ij} \quad \text{where } N = \sum_{ij} 1 \quad (9)$$

such that in the reference state, the specificity of each contact is lost, and the remaining energetic contribution in a model with a 5 Å interaction radius simply arises due to the fact that desolvation has taken place. This choice of reference state has the simple interpretation that formation of those contacts that are observed in the database more frequently than average is favored, whereas formation of those contacts that are observed rarely is penalized.

This choice of reference state also has effectively unrestricted spatial sampling of the ligand with respect to the protein and vice versa. In essence, it has no notion of chemical structure. And, since the specificity of each contact is lost, the only entropic contribution is precisely the entropy due to configurational freedom. Hence, subtracting \bar{g} from g_{ij} accounts for the entropic effect of restricted sampling as well as the energetic effect of desolvation.

This model is used to score candidate structures by an evaluation of the total binding free energy

$$G = \sum_{ij} g_{ij} \Delta_{ij} \quad (10)$$

where Δ_{ij} is zero unless i and j are within 5 Å of each other, in which case it is one. Thus, with this choice of model and reference state, G is an approximation to the complete change in free energy upon complex formation. Coarse graining has included entropic effects of solvation, and the reference state has provided the effects of solvation energy and configurational entropy.

One final aspect of the model is that the number of atom types is expanded to include some notion of the chemical personality of the various atoms. In other words, carbon atoms are broken into the categories of fatty carbons and polar carbons, and oxygen atoms are either charged, hydrogen bond donors, or hydrogen bond acceptors. Similarly nitrogen atoms and some other atoms and ions are included, such as sulfur, phosphorus, fluorine, calcium, and zinc. The model, together with the knowledge-based potential, is referred to as the design energy in this work.

Databases. Testing and application of SMoG has been subdivided into two parts: binding to sites on a protein surface and binding to sites in pockets. This subdivision is based on the observation that significantly different probabilities of interaction arise in each case, largely due to the different role of solvent in each situation which is reflected in the different contributions from solvation/desolvation terms. For the protein surface work, 17 complex structures were chosen: 1cmc 1dhi 1ela 1glq 1gmp 1hew 1nco 1nsc 1nsd 1pip 1sha 1sre 1tlm 2msb 2ohx 2sar 4dfr. These are all unique high resolution (≤ 2.0 Å RMSD) structures of non-peptide ligands bound to surface receptors. For the non-surface work, the training database included the following complexes (also ≤ 2.0 Å RMSD): 1art

lbed lbcx lbic lbit lbyb lcah lcam lcan lcao lcaz lchn lcll
lcmp lcoy lera lcrq lcsd lcsf lcsi lenc lerb lfel lfem lfen
lfdk lfkf lfkx lgca lgcd lhcb lhlh lhvi lhvk lhvl lhyt licm
licn linc lisc llcc llie llid llie llif llra llst lmdq lmf1 lmnq
lolib lpal lpbe lpbp lppf lppp lray lraz lsnm lsta lsty lswm
lsyd lthl ltng ltnh ltni ltnj ltnk ltnl ltpq ltrt 2aae 2acq 2acr
2acs 2acu 2che 2csc 2ctc 2cut 2fke 2mbp 2pal 2rnt 2tbs 2xis
3cla 3cts 3dfr 3gch 3pat 3rnt 4csc 4gch 4pal 4sga 5cts
5sga 5tim 6rnt 7rnt 821p 8est 8tlm 8xia 9est.

Monte Carlo Molecular Growth Algorithm. Directly in the binding region of the protein, simple organic molecules are generated by joining fragments with single bonds. Each step of the molecular growth proceeds as follows: two hydrogen atoms are selected—one from the fragment to be added and one from the structure as generated so far. The new fragment is placed such that the hydrogen atoms are displaced, and the atoms formerly bonded to those hydrogen atoms now form a single bond with a standard bond length. This procedure ensures that the new bond angles and bond lengths are reasonable approximations. Finally, the new functional group is oriented by torsional rotation about the new bond. Table 1 lists the fragments used in molecular growth.

In this manner, beginning with simple H_2 in the binding site, a molecule of any desired size can be generated, by continuing to add fragments. Notice that the growth is inherently branched because at each growth step any hydrogen atom on the present structure is a potential site of growth.

Each fragment that is placed is oriented by torsional rotation about the new bond in fixed increments (taken to be 60 degrees), and all those orientations that are not sterically hindered (*i.e.*: leading to atom pairs within 70% of the sum of their van der Waals' radii) are subject to energetic evaluation. That rotamer with the lowest energy is considered as a candidate for acceptance into the new molecule. This acceptance is determined by a metropolis Monte Carlo criterion which compares the new energy per atom with that before this growth step. Any decrease is accepted, and any increase is accepted with probability $\exp[-\Delta g/T]$ where $g = G/N$ is the free energy per atom, and T is an algorithmic temperature.

The preliminary selection of lowest allowed rotamer has two positive effects. First, it biases the molecule more quickly to low energy, since random selection of rotamers would lead to significantly more metropolis failures. Second, it is an indirect selection toward the tightest possible steric complementarity.

The Metropolis decision of acceptance or rejection of the new fragment is in place to allow the energy per atom to increase occasionally, as would need to be the case if the small molecule had grown into a tight steric region and had no other recourse but to grow into the solvent or some other unoccupied region, where it would interact only marginally with the protein.

Analysis of Growth Algorithm. In any implementation of an algorithm such as presented here, care needs to be taken in selecting the global parameters in the algorithm. These include the algorithm's temperature, the nearest approach allowed between atoms when assessing steric hindrance, and the angular increment in choosing the fragment rotamers. The nearest approach distance was taken to be 70% of the sum of the van der Waals radii of the atoms under consideration, since this gave good correlation with the nearest approach distances observed in the database. Sixty degree increments were chosen in part because finer increments gave rise to significantly more lengthy computation times, and because finer resolution was not consistent with the coarse grained potential. Selection of the optimal running temperature was made by observing the distribution of energies and computation time at different temperature. Figures 1 and 2 show the mean energy and

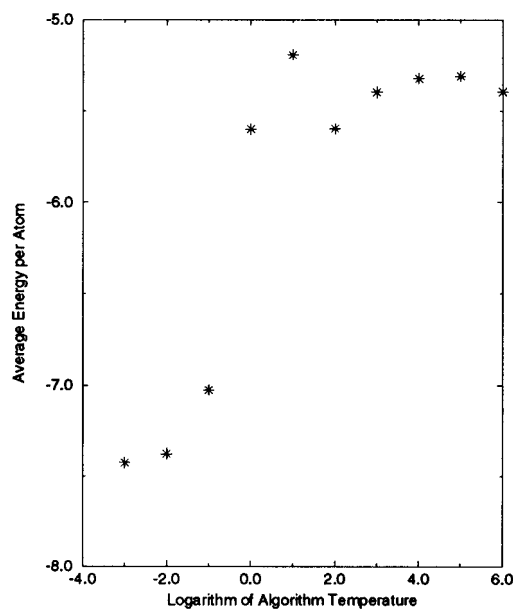


Figure 1. The average energy for ligands generated for 1sre at various algorithmic temperatures (log plot chosen for clarity of display only). As with all Monte Carlo algorithms, the algorithmic temperature defines how the algorithm responds to steps which increase the parameter being minimized. Higher temperature implies higher probability of acceptance. Here the effect of such a parameter on the final energies per heavy atom of the molecules generated by SMOG is shown. There is a sharp affect in the narrow range of temperatures near $T = 1$.

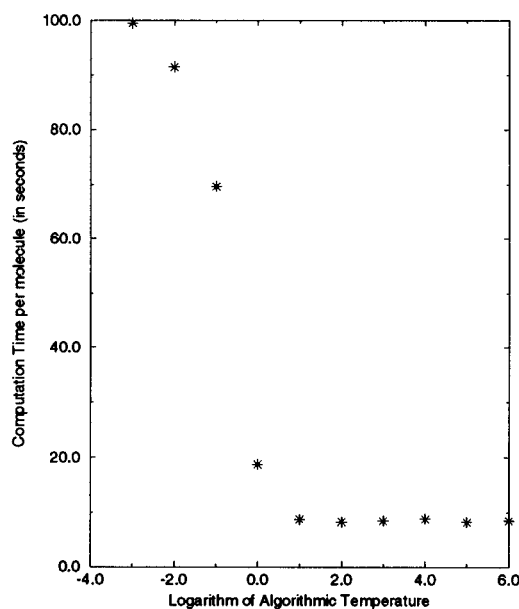


Figure 2. The average computation time for ligands designed for 1sre at various algorithmic temperatures (log plot chosen for clarity of display only). At the same temperature for which the average energy of the molecules rises, the algorithm becomes much more efficient. This results from the fact that a higher acceptance rate of molecular fragments implies quicker but less selective growth. However, the trade off in efficiency needs to be viewed with a pragmatic attitude in this situation, since the key parameter to optimize is the probability to generate extremely low energy molecules in a given time. Certainly, higher operating temperatures are preferred since the affect on computation time is more drastic than the affect on average energy per heavy atom.

computation time for generation of a thousand ligands to streptavidin, one of the surface proteins. Notice that there seem to be two regimes of operation of the algorithm, high bias and low bias. Because the optimal algorithmic temperature is the one that generates the largest number of low energy structures

per unit time, the low bias mode was selected for the balance of the work in this paper, namely $T = 10.0$.

The two modes can be understood from the point of view of the solvation energy and configurational entropy. There are some configurations whose free energy is favored because of the desolvation of lipophilic regions of the ligand and protein. However, the restrictions of the sampling of the space may make such configurations relatively improbable. Given a sufficiently low algorithmic temperature, the persistent algorithm will attain these configurations at the expense of time.

Under the operating conditions of 60 degree torsional increments, 70% van der Waals contact radius, and an algorithmic temperature of 10.0, each molecule of about 20 heavy atoms can be generated in a few seconds on a 100 MHz pentium computer running Linux.

Synopsis of Program. SMoG can be operated in several modes, which can be summarized as automatic, directed, or assisted. Automatic generation requires only the input of the protein structure and a coordinate used to specify the vicinity of the binding site, from which it proceeds to generate ligands with at least one atom within 5 Å of the specified coordinate. Directed mode is an interactive program that allows the user to specify which molecular fragments are selected and where they are attached. This mode allows the user to specify a specific molecule. Assisted growth begins with a user specified restart fragment but proceeds from that fragment automatically. This mode allows the user to incorporate a specific fragment into each molecule. The program also contains a conformational search facility, which performs a search in the space of interfragment torsion angles for the conformation with the lowest interaction energy.

Results

For the SMoG method to be proven effective, several requirements must be met. First the design potential must recognize native ligands (*i.e.*: ligands known to bind) as extremely low in free energy compared to an ensemble of generated molecules. In so doing, the scoring method reflects the fact that native ligands have a large negative binding free energy. Second, the algorithm must be able to generate some molecules with free energies comparable to native ligands in a reasonable amount of computation time. This demonstrates that SMoG can generate complexes with free energies comparable to a known ligand. Third, since the binding energy is a large component of the free energy, there must be some rough correlation between the design energy and an estimation of binding energy using an empirical force field such as CHARMM. Fourth, there needs to be evidence that the guiding knowledge-based potential can be relied upon to reproduce experimental binding free energies, in order to establish the knowledge-based potential. Finally, the molecules generated by SMoG must not only score well quantitatively but must be qualitatively appealing as well. The evidence that SMoG meets each of these requirements is given in the following sections.

Attaining and Discriminating True Ligands. Figures 3 and 4 demonstrate, for each of the complexes in the surface database save one, that the knowledge-based potential respects the native ligand (whose energy is marked as a dark stripe) as having extremely low energy. Moreover, molecules with a comparable energy are rare but attainable in reasonable computation time since approximately 5% of generated molecules are comparable to the native ligand in each example save one.

The exception is an example where the native ligand contains only four atoms. Most likely, the algorithm would only take one step, with little opportunity for biasing, or little need to compromise energy for steric freedom.

Quasi-Correlation with Empirical Binding Energies. In order to examine rough correlation between the design energy approximation to free energy and an empirical estimation of binding energy, the protein streptavidin (Isre) was chosen because its native ligand scored exceptionally strongly in design energy and the native ligand and protein were both rather small, making subsequent calculations with CHARMM more efficient. Fifty of the lowest energy molecules generated with SMoG were minimized to convergence in the binding site of streptavidin. Figure 5 shows the correlation of the CHARMM interaction energy with the design potential. We are not seeking to demonstrate a one-to-one correspondence with CHARMM but rather to show that those molecules with low design energy also have low empirical energy. Also, there is a rough correlation between the two scoring methods, which is to be expected as the design energy, as an estimate of binding free energy contains a large contribution from binding energy. It should be noted here that the empirical energy is a vacuum enthalpy estimate and, therefore, provides unreliable estimates of the solvent effects such as hydrophobic interaction. Thus the scatter in this figure results from the entropic and solvent energy factors in the free energy. This result also demonstrates that the SMoG algorithm and statistical potential are able to generate ligands that are predicted to have binding energies as strong as the native ligand.

Because the SMoG estimation of free energy is an estimate, rather than an accurate determination, a recommended protocol for the screening of lead candidates is to perform empirical estimations of the binding energy and select as candidates for further testing (be it experimental or computational) those that score best in both binding energy and design free energy. These are the candidates below the shaded line in Figure 5. Indeed, examination of the structures of these molecules in complex with streptavidin showed the presence of good steric complementarity, several hydrogen bonds, and association of lipophilic moieties: the qualitative features desirable in ligand design.

Correlation with Experimental Binding Free Energies. In order to test the correlation between experimental binding free energies and the SMoG design procedure, SMoG was applied to the three protein-ligand complex systems for which structural and binding information has been published and is readily available. These examples include purine nucleoside phosphorylase (PNP), Src SH3 domain specificity pocket (SH3 domain), and human immunodeficiency virus-1 protease (HIV). Each case will be presented in turn.

Purine Nucleoside Phosphorylase. Guanine based ligands that have been designed, synthesized, and assayed for purine nucleoside phosphorylase (PNP).⁴⁻⁹ In these publications, the authors present their rationale for synthesizing the ligands that they tested, which rests on computer models of the ligands, each of which adopts a binding mode defined in part by the coordinates of guanine in the crystal structure 1ulb and in part by a combination of conformational search and energy minimization with an empirical force field.

(4) Tuttle, J. V.; Kernitzky, T. A. *J. Biol. Chem.* **1984**, 259, 4065-4069.

(5) Ealick, S. E.; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Montgomery, J. A.; Secrist, J. A. *PNAS* **1991**, 88, 11540-11544.

(6) Montgomery, J. A.; Niwas, S.; Rose, J. D.; Secrist, J. A.; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Ealick, S. E. *J. Med. Chem.* **1993**, 36, 55-69.

(7) Secrist, J. A.; Niwas, S.; Rose, J. D.; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Guida, W. C.; Ealick, S. E.; Montgomery, J. A. *J. Med. Chem.* **1993**, 36, 1847-1854.

(8) Erion, M. D.; Niwas, S.; Rose, J. D.; Subramanian, A.; Allen, M.; Secrist, J. A.; Babu, Y. S.; Bugg, C. E.; Guida, W. C.; Ealick, S. E.; Montgomery, J. A. *J. Med. Chem.* **1993**, 36, 3771-3783.

(9) Guida, W. C.; Elliott, R. D.; Thomas, H. J.; Secrist, J. A.; Babu, Y. S.; Bugg, C. E.; Erion, M. D.; Ealick, S. E.; Montgomery, J. A. *J. Med. Chem.* **1994**, 37, 1109-1114.

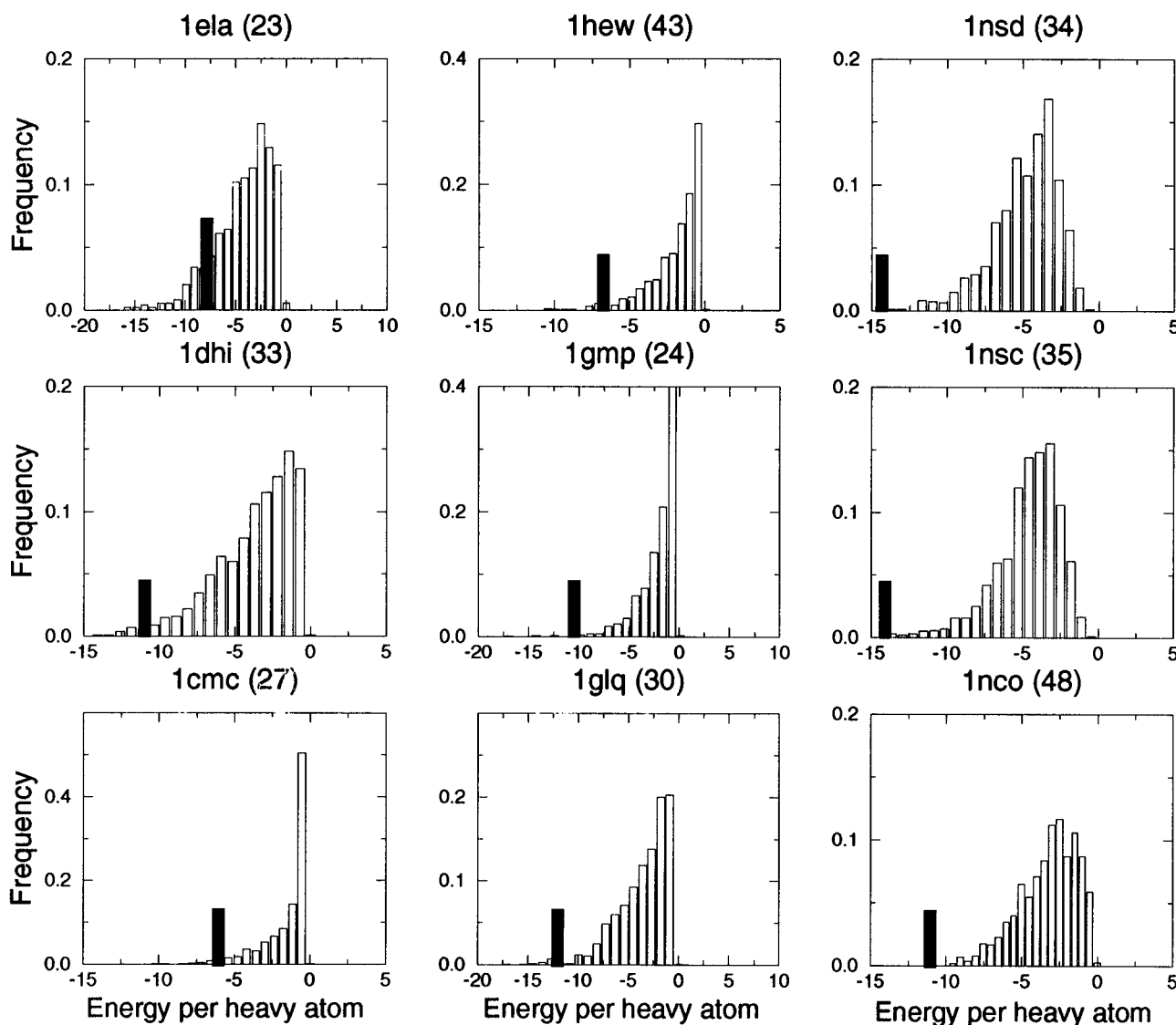


Figure 3. The distribution of energy for the design of 1000 molecules of the same size as the native ligand (sizes shown in brackets) for the first nine complexes in the surface database. Notice that the energy of the native ligands, shown in black, are always in the extreme tail of the distribution. The differences in the ranges of the energy per atom reflect the differing character of the binding sites and the various sizes of the small molecules. Notice, however, that regardless of these two factors, the positioning of the native ligand's energy in the distribution is the same in each case. This implies both that the native ligand has an extremely rare SMOG energy and that the algorithm is able to generate comparable ligands. Both of these factors support the hypothesis that the course-grained potential reflects the binding free energy of the complexes.

Accordingly, each of the molecules in Table 2 was built interactively with SMOG (directed mode), and the lowest energy conformation was found with SMOG's conformational search facility. In this sense, the molecules were tested as if they had been generated by SMOG's *de novo* growth algorithm. That is to say that, given enough time, SMOG would have generated these molecules and the corresponding conformations. However, undirected generation of these exact ligands is a highly improbable event. The result is that we are testing a set of the molecules generated by SMOG for correlation between free energies (taken as the logarithm of the binding constants or IC_{50} measurements) and SMOG's knowledge-based potential. This approach was used in the SH3 domain and HIV cases as well.

The PNP binding site, however, contains a pocket for phosphate as well as a nucleoside, and the binding constant (K_i or IC_{50} depending on the affinity) of each of the ligands was determined at two different concentrations of phosphate (1 and 50 mM), and some molecules showed high sensitivity to the phosphate concentration. Because the SMOG conformational search and estimation of the binding free energy did not account for the presence or absence of the phosphate (indeed it is unclear how to do that without introducing untestable hypotheses), one

can only expect that SMOG's score would correlate with experimental measurement for those ligands which were insensitive to the phosphate concentration and at the lower concentration. As Figure 6 shows, this is indeed the case: the highly sensitive molecules show no correlation with SMOG, whereas the others show very strong correlation. The significance of these two observations is taken up in the discussion.

SH3 Domain. In a separate system, the specificity pocket of SH3 domains,^{10–12} a similar test was performed. The coordinates of one ligand was provided to us by Sibo Feng and Stuart Schreiber, which represented a superset of several of the other experimental ligands. By trimming this structure down, several ligands were prepared (see Table 3). The remaining ligand, which was structurally independent, was generated as described for the PNP ligands. Figure 7 shows the correlation of experimental binding constant and SMOG's estimation of the free energy of binding.

(10) Chen, J. K.; Lane, W. S.; Brauer, A. W.; Tanaka, A.; Schreiber, S. *J. Am. Chem. Soc.* **1993**, *115*, 12591–12592.

(11) Feng, S.; Chen, J. K.; Yu, H.; Simon, J. A.; Schreiber, S. L. *Science* **1994**, *266*, 1241–1247.

(12) Combs, A. P.; Kapoor, T. M.; Feng, S.; Chen, J. K.; Daude-Snow, L. F.; Schreiber, S. *J. Am. Chem. Soc.* **1996**, *118*, 287–288.

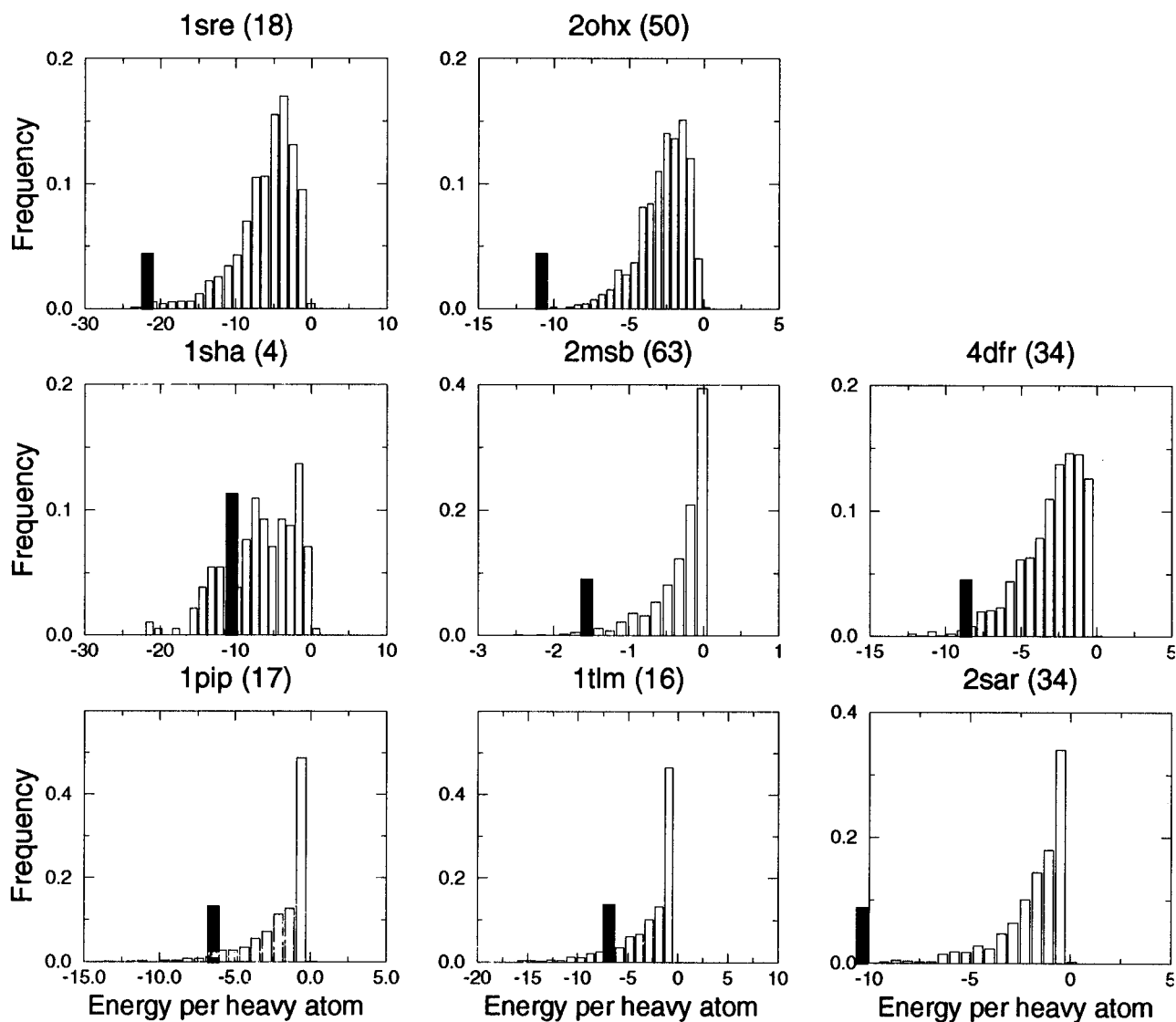


Figure 4. The distribution of energy for the design of 1000 molecules of the same size as the native ligand (sizes shown in brackets) for the last eight complexes in the surface database. With one exception, the energy of the native ligands, shown in black, are always in the extreme tail of the distribution. The differences in the ranges of the energy per atom reflect the differing character of the binding sites and the various sizes of the small molecules. Notice, however, that regardless of these two factors, the positioning of the native ligand's energy in the distribution is the same in each case. These examples provide further support of the course-grained potential.

HIV-1 Protease. HIV-1 protease has been the target of very much structure-based drug design effort, and as such there is a wealth of literature on the subject. However, in choosing a system of ligands for proofing the correlation between SMoG's course-grained potential and experimentally determined binding free energies, several considerations need to be applied. First the experimental determinations have to have been performed under identical conditions among the members in the system. Secondly the binding constants must span a wide range. Thirdly binding mode coordinates must either be published or attainable via conformational search. Fourth, the molecules must be structurally diverse (SMoG is not an effective lead optimization tool—see the Discussion section) and yet of roughly the same molecular weight. The system we have studied^{13–14} is presented in Table 4, and the results are plotted in Figure 8.

(13) Abdel-Meguid, S. S.; Metcalf, B. W.; Carr, T. J.; Demarsh, P.; DesJarlais, R. L.; Fisher, S.; Green, D. W.; Ivanoff, L.; Lambert, L.; Murthy, K. H. M.; Petteway, S. R., Jr.; Pitts, W. J.; Tomaszek, T. A., Jr.; Winborne, E.; Zhao, B.; Dreyer, G. B.; Meek, T. D. *Biochemistry* **1994**, *33*, 11671–11677.

(14) Thompson, S. K.; Murthy, K. H. M.; Zhaong, B.; Winborne, E.; Green, D. W.; Fisher, S. M.; DesJarlais, R. L.; Tomaszek, T. A., Jr.; Meek, T. D.; Gleason, J. G.; Abdel-Meguid, S. S. *J. Med. Chem.* **1994**, *37*, 3100–3107.

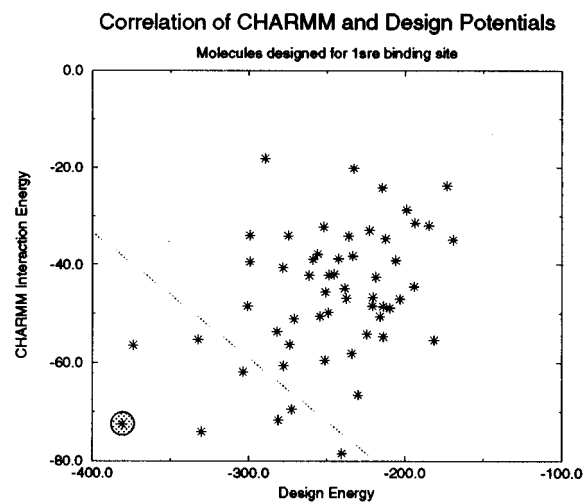
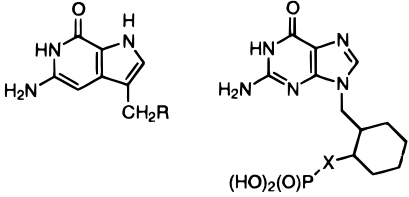


Figure 5. The correlation between the energies of designed ligands as determined by the knowledge-based design potential and the empirical CHARMM potential after complete minimization. Notice the placement of the native ligand (circled) and the proximity of several other molecules. Those below the arbitrarily drawn gray line are good candidates for binding.

Table 2. The PNP Ligands Tested by SMOG's Course-Grained Potential and Conformational Search Facility^a


R	phosphate sensitivity	K_i or IC_{50} (μM) (1 mM phosphate)	SMoG energy per heavy atom
2-hydroxyphenyl	low	0.27	-18.1
2-tetrahydrofuranyl	low	0.07	-16.2
2-tetrahydrothienyl	high	0.011	-16.6
2-thienylmethyl	low	0.021	-16.6
3-methoxyphenyl	low	0.082	-18.1
3-methylcyclohexyl	high	0.025	-18.0
3-thienylmethyl	low	0.025	-15.8
3-trifluoromethylcyclohexyl	high	0.025	-13.2
3-trifluoromethylphenyl	low	0.036	-12.3
4-hydroxyphenyl	low	0.26	-18.7
cycloheptyl	high	0.03	-17.1
cyclohexyl (no methylene)	high	1.3	-17.0
cyclohexyl	high	0.047	-17.4
cyclopentyl	high	0.029	-18.0
methylphenyl	low	0.057	-19.4
phenyl	low	0.051	-18.7
pyridin-3-yl	low	0.025	-18.5

X	phosphate sensitivity	K_i or IC_{50} (μM) (1 mM phosphate)	SMoG energy per heavy atom
-(CH ₂) ₂ -	low	0.035	-17.8
-(CH ₂) ₃ -	high	0.62	-18.8
-O(CH ₂) ₂ -	high	1.00	-18.9
GMP	low	530	-14.1
GDP	low	360	-13.9
GTP	low	490	-14.7
dGMP	low	300	-14.4
dGDP	low	37	-14.9
dGTP	high	32	-14.4
acyclovir	low	100	-15.8
acyclovirMP	low	6.6	-14.4
acyclovirDP	high	0.009	-14.4
acyclovirTP	high	0.31	-14.4

^a Each molecule contains a guanine or 9-deazaguanine fragment, which was held fixed at the coordinates in the 1ulb crystal structure of guanine. The binding mode of the balance of the structure was determined by conformational search on the potential surface provided by SMOG's course-grained potential. Those molecules marked as having low phosphate sensitivity are those whose binding constant changes by a factor of 15 or less upon increase of the concentration of phosphate to 50 mM. The highly sensitive molecules are affected in some instances by a factor of 140.

Table 5 summarizes the overall correlation findings quantitatively.

Example of de Novo Design—CD4. The CD4 protein is an immunoglobulin-family transmembrane coreceptor expressed in the helper T-cells. It participates in contact between the T-cells and antigen-presenting cells by binding to the nonpolymorphic part of the class II major histocompatibility complex (MHC-II) protein, which is followed by the activation of the bound Lck kinase activating the T-cell.

The human immunodeficiency virus (HIV) disrupts the immune response mechanism by binding to CD4, penetrating into the T-cells, and killing them. Therefore in order to prevent HIV binding and subsequent action, the effort to find an inhibitor to the binding between gp120 of HIV and CD4 is ongoing.

Figure 9a shows the chemical structure of a candidate ligand for the binding site in the vicinity of Phe 43 of CD4 (Figure 9b). *De novo* growth with SMOG presented several molecular

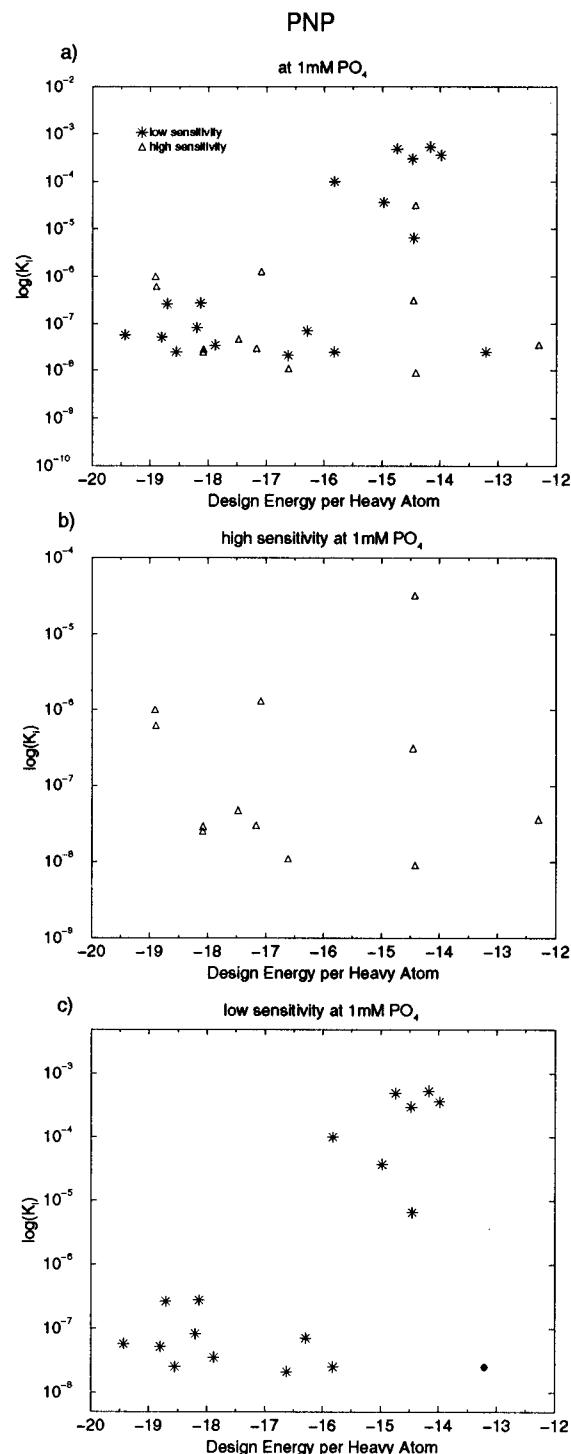
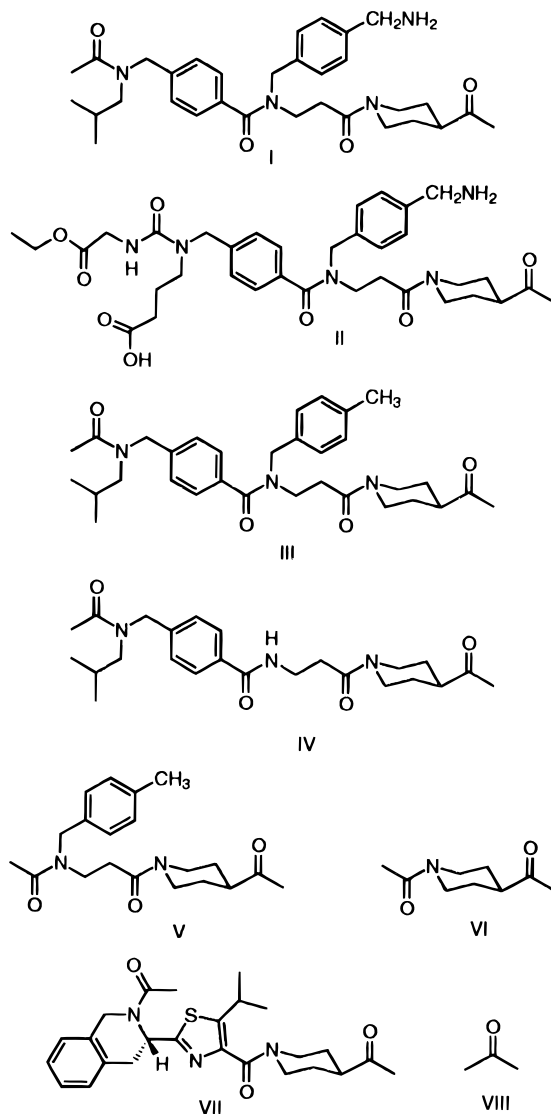


Figure 6. Measuring the correlation of SMOG's course-grained potential and experimental binding constants in a series of purine nucleoside phosphorylase inhibitors. Binding constants are plotted on the log scale since the logarithm of the binding constant is proportional to the experimental binding free energy. a) All the molecules are from Table 2, showing no apparent significant correlation. However, classification of the ligands into those whose binding is highly sensitive to the phosphate concentration (b) and those that are relatively insensitive (c) demonstrates that the noise in plot (a) is largely due to the highly sensitive ligands. Indeed, (b) shows absolutely no correlation, whereas (c) shows a significant correlation. One outlier in graph (c), in the lower right, is a molecule with three fluorine atoms. Since fluorine appears only seldom in the database of crystal structures, the interaction parameters for fluorine are ill-defined.

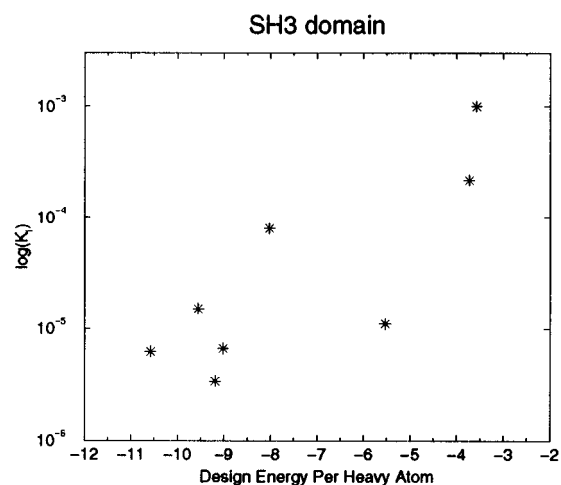
scaffolds which each contained positive features that attributed to their low binding free energy estimate, such as a cluster of three hydrogen bonds from a sugar-like ring shown at the bottom

Table 3. The Src SH3 Domain Specificity Pocket Ligands Tested by SMOG's Course-Grained Potential and Conformational Search Facility^a

ligand	K_i (μM)	SMoG energy per heavy atom
I	3.4	-9.1
II	6.6	-9.0
III	6.2	-10.5
IV	80	-8.0
V	15	-9.5
VI	220	-3.7
VII	11	-5.5
VIII	1000	-3.5

^a We were graciously provided the NMR structure for molecule I complexed with SH3 by Sibö Feng and Stuart Schreiber from which we were able to generate binding modes for molecules IV, V, VI, and VIII. The binding mode of molecules II and III were determined by conformational search using molecule I as a template. The binding mode of molecule VII was determined by conformational search on the potential surface provided by SMOG's course-grained potential, using the carbonyl group from molecule I as a fixed fragment (this group provides the link to the peptide biasing element, which was not included in these structures.)

of Figure 9a and a partial π -stacking with Phe 43. By manual addition of a methylene group between the sugar's 4' ester linkage and the pyridine fragment and subsequent minimization of the structure with CHARMM, the π -stack was improved, and the resulting geometry suggested the formation of the seven-membered fused ring bridge to increase the rigidity of the molecule and lock in the relative orientation of the pyridine ring and the hydrogen-bonding groups. Finally, the hydroxyl

**Figure 7.** Measuring the correlation of SMOG's course-grained potential and experimental binding constants in a series of ligands for the specificity pocket of Src SH3 domain. Binding constants are plotted on the log scale since the logarithm of the binding constant is proportional to the experimental binding free energy. As in the case of the PNP ligands with low sensitivity to phosphate concentration, there is considerable correlation.

group on the seven-membered ring was added to take advantage of a potential hydrogen bond which also suggested itself. The resulting molecule and the interactions it makes with the protein are shown in Figure 9.

Though there was considerable manual intervention in arriving at this specific ligand candidate, SMOG, in an unbiased design, suggested the key molecular fragments and provided molecules which displayed these fragments in the relevant orientation, thereby solving the bulk of the combinatorial problem in lead design.

Discussion

Before we discuss our new contributions to the field, it is appropriate to review the state of the art. With the exception of the MCSS based approaches, each of the following treats the protein rigidly. In each case, the overwhelming number of candidates is trimmed down significantly by application of screening to a large but manageable database or by trimming the search tree through molecular generation algorithms that strive to incorporate specific features that were found in initialization stages.

DOCK.¹⁵⁻¹⁷ This is predominately a geometric method wherein the Connolly surface of the receptor¹⁸⁻²⁰ is mapped onto a negative image. This negative image is used as the search target for similarity with molecules in a library. Scoring is either done with qualitative assessment of potential hydrogen bonding and charge pairing or with estimation of interaction energy with an empirical potential. DOCK is particularly able to find the correct binding mode of ligands known to bind. In this regard, however, it is limited by the assumption of fixed geometry of the ligand as well as the extent of the library of potential candidates. As the method relies on libraries of complete molecules, it is unable to suggest novel structures.

GRID.²¹ This seminal work forms the seed of many of the algorithms and approaches that have come since. Using an

(15) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1982**, *161*, 269-288.

(16) Desjarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. *J. Med. Chem.* **1986**, *29*, 2149-2153.

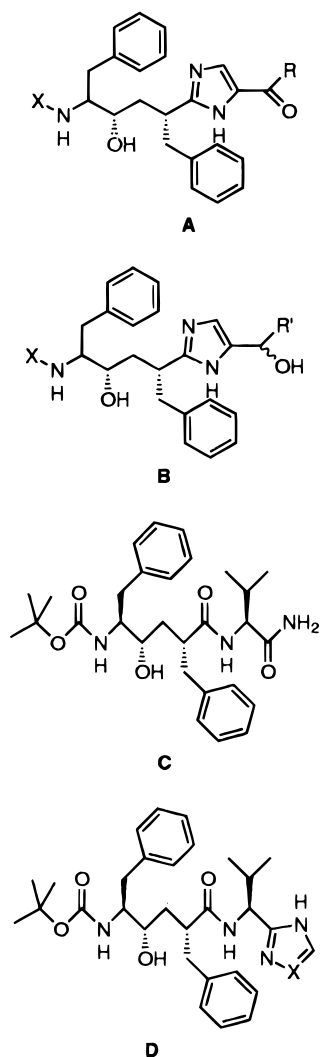
(17) Shoichet, B. K.; Kuntz, I. D. *J. Mol. Biol.* **1991**, *221*, 327-346.

(18) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379-400.

(19) Connolly, M. L. *J. Appl. Cryst.* **1983a**, *16*, 548-558.

(20) Connolly, M. L. *Science* **1983b**, *221*, 709-713.

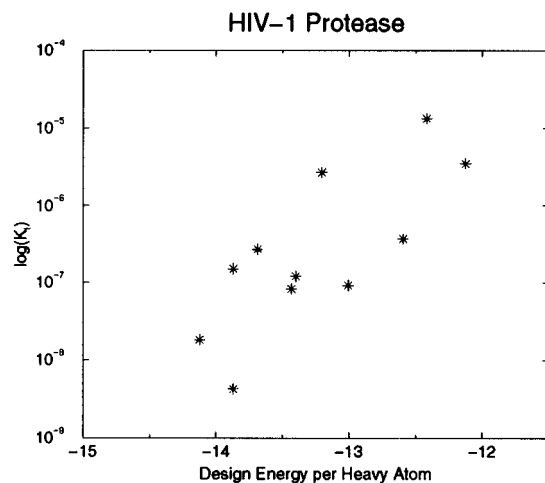
(21) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849-857.

Table 4. HIV-1 Protease Ligands Tested by SMOG's Course-Grained Potential and Conformational Search Facility^a

molecule	X	R (or R')	K_i (nM)	SMoG energy per heavy atom
A	Boc	H	3500	-12.1
A	Boc	Me	370	-12.5
A	Boc	Et	92	-13.0
A	Boc	<i>n</i> -Pr	150	-13.8
A	Boc	<i>i</i> -Pr	83	-13.4
A	Boc	CM_2CHCH_2	270	-13.6
B	Boc	Me(R)	13300	-12.3
B	Boc	Me(S)	13300	-12.4
B	Boc	<i>i</i> -Pr(R)	2700	-13.3
B	Boc	<i>i</i> -Pr(S)	2700	-13.0
C			1.4	-13.3
D	CH		18	-14.1
D	N		4.2	-13.8

^a As in the PNP case, these molecules share common structural motifs, so that by using the crystal structures (1hps and 1sbg) to define the coordinates of these motifs and using conformational search on the balance of each molecule, the binding mode of each ligand was determined. Also, the structurally specific waters in the binding site were included as part of the protein.

empirical hydrogen-bonding interaction potential and spherical representations of functional groups, GRID generates affinity contours for various molecular fragments which identifies regions of high and low affinity. These contours can be used to guide chemical intuition or as input to several analysis programs. GRID is limited by its representation of the fragments, which does not allow prediction of fragment orientation.

**Figure 8.** Measuring the correlation of SMOG's course-grained potential and experimental binding constants in a series of ligands for the specificity pocket of Src SH3 domain. Binding constants are plotted on the log scale since the logarithm of the binding constant is proportional to the experimental binding free energy. As in the case of the PNP ligands with low sensitivity to phosphate concentration, there is considerable correlation.**Table 5.** Summary of Correlation Data^a

system	correlation coefficient	no. of points	probability of random occurrence
PNP ^b	0.80	17	0.002
SH3	0.81	8	0.110
HIV	0.77	11	0.050

^a Here are presented the correlation coefficients for each of the preceding ligand systems. Note that in each case there is significant correlation. The probability of random occurrence is the probability that a random selection of the same number of points would have the given correlation constant. In other words, the confidence that the observed correlations are systematic (and not the result of sparse sampling) are 99.8%, 88.9%, and 95.0%. Taken together, these data imply that the confidence in correlation between SMOG's course-grained potential and the experimental binding free energy is established.]^b This correlation applies to the low sensitivity data only.

GROW.²² By joining peptide fragments from an extensive conformational library, this method generates peptide ligands in a sequential molecular growth algorithm. Scoring includes empirical interaction energy and internal energy as well as surface area terms to approximate solvent effects.

LUDI.^{23,24} According to simple, qualitative rules, favorable sites are located for various functional groups which are then joined together with linker fragments. Beyond ensuring that steric clashes are avoided, no scoring of the new candidates is performed. LUDI also allows the use of precalculated interaction sites as produced by GRID. This method is exceptionally quick and, therefore, can be used interactively.

CLIX.²⁵ As an enhancement over early versions of DOCK, CLIX provides a screening of a structural library with respect to patterns of functional groups as determined by GRID. Beyond assuring no steric clashes, this method scores the candidates by summing the energetic contributions (as determined by GRID) of each functional group that matches the search template.

MCSS-HOOK-DLD.²⁶⁻²⁹ These novel approaches involve a sophisticated, dynamic treatment of the protein binding site,

(22) Moon, J. B. and Howe, W. J. *Proteins* **1991**, *11*, 314-328.(23) Böhm, H.-J. *J. Comput-Aided Mol. Design* **1992a**, *6*, 61-78.(24) Böhm, H.-J. *J. Comput-Aided Mol. Design* **1992b**, *6*, 593-606.(25) Lawrence, M. C.; Davis, P. C. *Proteins* **1992**, *12*, 31-41.(26) Miranker, A.; Karplus, M. *Proteins* **1991**, *11*, 29.(27) Caflisch, A.; Miranker, A.; Karplus, M. *J. Med. Chem.* **1993**, *36*, 2142-2167.

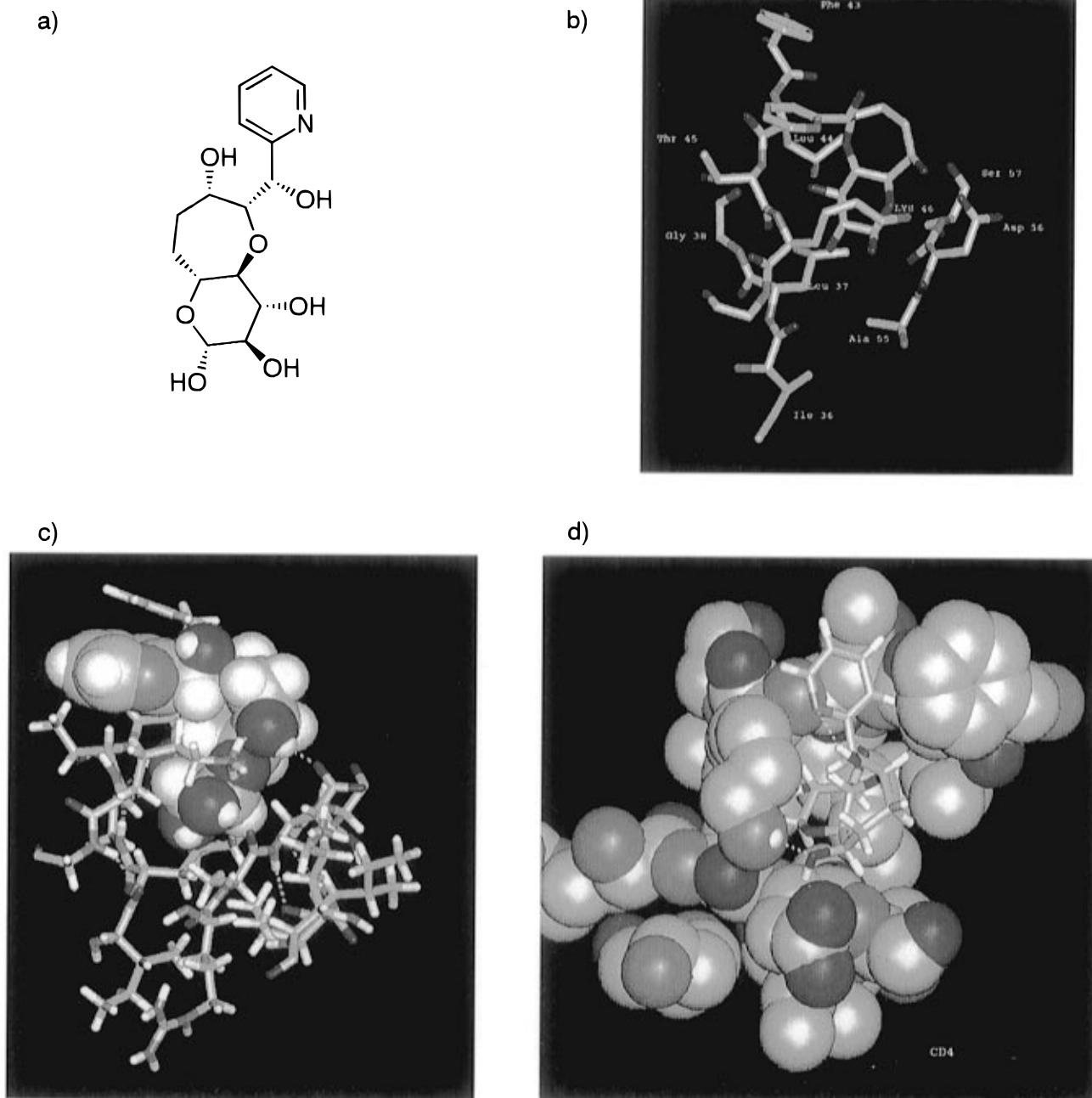


Figure 9. A candidate ligand for the Phe 43 binding pocket of CD4. This molecule is able to form five hydrogen bonds (four intermolecular and one intramolecular) as well as a significant π -stack with the benzene ring of Phe 43. (a) Molecular structure of the candidate: note the rigid structure. (b) Licorice diagram of the ligand in the binding site showing the residues with which a strong ligand should make interactions. (c) The ligand shown as a space filling model. Notice the π -stacking with Phe 43. (d) Another view, this time with the protein as a space-filling model.

which locates favorable interaction sites for molecular fragments by performing a multiple copy simultaneous search. In such a search, the protein is subject to the average potential field of the ligands using the CHARMM empirical force field. The resulting interaction sites, unlike with GRID, contain orientation information and can be linked together with bonding force fields and linker sp^3 and sp^2 carbon atoms (DLD, dynamic ligand design) or molecular fragments in a database (HOOK). Although the MCSS-based approaches rely essentially on binding energy calculations for scoring, they are the first step toward taking all of the relevant degrees of freedom into account in the ligand design process, since the ligands and the protein are flexible. The unfortunate aspect of this sophisticated approach is the large amount of computation

required, several days preparation time on a modern workstation followed by approximately an hour of computation for each ligand candidate.

SMoG. As has been shown, the knowledge-based potential discriminates very clearly between those molecules which are likely to bind well and those that are not. In the case of the surface proteins, all the native ligands (save one) were found to lie at the tail of the distribution of free energies of molecules that SMoG was able to derive (which themselves were already the result of a minimization in the form of a biased pruning of the search tree). This was corroborated with the enthalpic comparison performed using a well-accepted empirical force field.

Also, as shown in the studies of the PNP, SH3 domain, and HIV-1 protease ligands, the course-grained knowledge-based potential correlates very strongly with the experimental binding free energies insofar as it is reasonable to expect such a

(28) Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. *Proteins* **1994**, *19*, 199–221.

(29) Miranker, A.; Karplus, M. *Proteins* **1995**, *23*, 472–490.

correlation. Indeed, the PNP ligands whose binding constants are sensitive to the concentration of phosphate provide a control experiment against which to measure the correlations in the other cases. One observes, as one should, no apparent correlation between the design energy and the experimental free energy in a situation for which there is no *a priori* relationship between the two. Also, it should be made clear that SMOG is not yet at an appropriate level of accuracy for performing lead optimization. This is evidenced in Figure 6c, which shows that though SMOG is readily able to distinguish micromolar binders from nanomolar binders, it is not able to discriminate among those ligands that bind in the submicromolar range. This latter range is precisely the province of lead optimization methods, including combinatorial chemistry, structure–activity studies (SAR), isosteric substitution, and medicinal chemistry. This distinction also arose in studying the HIV-1 protease inhibitors: those molecules which differed by subtle variation in one functional group scored similarly with SMOG, though they had radically different binding constants. Again, these small variations are the province of lead optimization, rather than lead discovery. We are currently pursuing enhancements to the interaction potential through the addition of specific interaction terms for hydrogen bond formation and salt bridges and expect to publish those results as the third paper in this series. Perhaps at that stage, we will be able to apply SMOG's scoring function in lead optimization studies and SAR analysis.

The two systems HIV and PNP can be combined to one correlation plot to obtain a very high overall correlation with slope of one ($r = 0.875$, $N = 30$, $P = 1.38 \times 10^{-8}$) leading to an overall relation between SMOG's free energy estimate ΔG_S and the experimental free energy ΔG_E :

$$\Delta G_E = \Delta G_S - 10.2 \quad (11)$$

However, the SH3 case does not fit into this scheme, but rather relates the two variables with a slope of 2.3 and an intercept of +3. The fundamental difference between these two classes of examples is that the SH3 case involves a series of surface binding ligands, whereas the other two enzymes bind their ligands in internal pockets. For ligands completely surrounded by protein, the number of intermolecular contacts (*i.e.*, protein atoms within 5.0 Å of the ligand atoms) is larger than the surface binding situation, the ratio of the numbers depending largely on the geometry of the binding site. Because SMOG's score is dependent on the number of contacts, the slope of the free energy prediction line will change from protein to protein; however, the relative values of the scores will be meaningful in all cases.

The somewhat surprising success of SMOG's simple interaction representation and non-empirical potential lies in the very nature of a knowledge-based energy applied to a coarse-grained model. In fact, by choosing the radius of interaction to be somewhat larger than intuition, we have subsumed much detail into our simple matrix of interaction free energies, g_{ij} . Because of the relation between SMOG's design energy and experimental free energies of binding, SMOG may provide a much needed tool that combines geometric fit with chemical intuition into a simple, quick, quantitative, predictive scheme. As such, SMOG may be useful in the development of novel lead compounds for systematic study and improvement in the pharmaceutical industry. This method is clearly able to discriminate between potential ligands that have a high probability of binding and

those that do not and is also capable of generating the favorable candidates quickly. Its potential applications range from computational brainstorming through explicit *de novo* design efforts. As a brainstorming tool, the molecules that SMOG produces can serve as a guided tour of a binding site, allowing one to visualize the possibilities for binding modes, specific interactions, and specific functional groups, through chemically viable molecules and fragments. As a companion to combinatorial chemistry efforts, SMOG's output may seed the inclusion of novel compounds into libraries. Furthermore, the program allows explicit inclusion of the tethering fragments and orientation of the novel molecule.

In the example of a CD4 lead candidate, SMOG was used both to explore the binding site and to arrive at a specific molecule which is rich in qualitative detail and scores very strongly relative to other molecules SMOG generated. As in the analysis of the known surface-binding ligands in Figures 3 and 4, we have strong reason to believe that this candidate will successfully bind to the CD4 binding site. The second paper in this series will discuss several examples of ligand candidates designed with SMOG and as well as highlighting the flexibility of ligand design with SMOG will provide a general methodology for developing novel molecules with a high propensity to bind to their targets.

In the role for which it has been designed, SMOG provides several advantages over other popular design methods. These include simple efficiency (each molecule taking just seconds on a personal computer), generating and evaluating whole molecules rather than separate fragments which later need to be linked, and, most importantly, documented correlation between the scoring method and free energies of binding.

SMoG's limitations include those implied in the simple methods with which chemical geometry is handled: interfragment bond lengths and angles are all assumed to be standard and unvarying; the protein structure is considered fixed; and steric repulsions are either on or off, depending on a simple distance test. Other limitations are implementation dependent, and the program has been designed to allow flexibility in the choice of operating conditions. For example, smaller angle steps can be chosen to perform calculations more carefully, lower temperatures can be chosen, and the fragment library can be expanded.

Of course, as is the case with any design method, the crucial test of SMOG's merit will include the synthesis and measurement of the binding constant of a candidate ligand that was the direct result of SMOG design. It is our goal to pursue this line of development vigorously.

Finally, the fact that SMOG's essential elements have been confirmed indicates that knowledge-based potentials and intermediate models of protein-ligand interactions are now a viable option for the study of many aspects of the binding problem which have until now been computationally foreboding.

Acknowledgment. The authors are grateful to the Packard Foundation and NSERC (Canada) for funding this research. We thank Fred Cohen for making us aware of the PNP system of ligands and Sibio Feng and Stuart Schreiber for the coordinates of one of the SH3 ligands. The hard work of Jun Shimada in the study of PNP, SH3, and HIV-1 protease systems is also acknowledged.

JA960751U