VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands

Richard D. Head, Mark L. Smythe,[†] Tudor I. Oprea,[‡] Chris L. Waller,[§] Stuart M. Green,[⊥] and Garland R. Marshall*

Contribution from the Center for Molecular Design, Washington University, St. Louis, Missouri 63110

Received November 20, 1995. Revised Manuscript Received February 16, 1996[⊗]

Abstract: VALIDATE is a hybrid approach to predict the binding affinity of novel ligands for receptors of known three-dimensional structure. This approach calculates physicochemical properties of the ligand and the receptor−ligand complex to estimate the free energy of binding. The enthalpy of binding is calculated by molecular mechanics while properties such as complementary hydrophobic surface area are used to estimate the entropy of binding through heuristics. A diverse training set of 51 crystalline complexes was assembled, and their relevant physicochemical properties were computed. These properties were analyzed by partial least squares (PLS) statistics, or neural network analysis (SONNIC), to generate models for the general prediction of the affinity of ligands with receptors of known three-dimensional structure. The ability of the model to predict the affinity of novel complexes not included in the training set was demonstrated with three independent test sets: 14 complexes of known three-dimensional structure including 3 DNA complexes, a class of compound not included in the training set, 13 HIV protease inhibitors fit to HIV-1 protease, and 11 thermolysin inhibitors fit to thermolysin.

Introduction

Advances in molecular biology, X-ray crystallography, and NMR spectroscopy are providing three-dimensional structures of potential therapeutic targets at atomic resolution at an increasing rate. Coupled with the vast increases in accessible computing power, these advances have ushered in a new era of rational drug design. Several structure-based drug design packages now exist[1−4] to assist in the design of novel ligands to interact with known three-dimensional active sites. A common strategy for structure-based drug design is to first search the known binding sites for positions that interact with different functional groups.[5,6] The next step is to connect the functional groups given their three-dimensional positions to form molecules that are candidates for synthesis.[7,8] The last step is the prediction of the affinity for the receptor of the designed ligands for synthetic prioritization.

One approach to predicting ligand affinity focuses on directly calculating the thermodynamics of forming a ligand−receptor complex. For a comprehensive review of the complexities involved and the various approaches, see the review by Ajay and Murcko[9] and the references therein. Williams[10−15] has used a vancomycin−peptide complex as an experimental system in which to evaluate the various contributions to binding affinity and produced the following relationship:

$$\Delta G_{\text{binding}} = \Delta G_{(\text{trans} + \text{rot})} + \Delta G_{\text{rotors}} + \Delta H_{\text{conform}} + \\ \sum \Delta G_i + \Delta G_{\text{vdW}} + \Delta G_{\text{H}}$$

$\Delta G_{(\text{trans} + \text{rot})}$ = the free energy associated with translational and rotational freedom of the ligand. This has an adverse effect on binding of 50−70 kJ/mol (12−17 kcal/mol) at room temperature for ligands of 100−300 Da assuming complete loss of translational and rotational freedom relative to the receptor. ΔG_{rotors} = the free energy associated with the number of rotational degrees of freedom frozen. This is 5−6 kJ/mol (1.2−1.6 kcal/mol) per rotatable bond, assuming complete loss of rotational freedom. $\Delta H_{\text{conform}}$ = the strain energy introduced by complex formation (deformation in bond lengths, bond angles, torsional angles, etc. from solution states). $\sum \Delta G_i$ = the sum of interaction free energies between polar groups. ΔG_{vdW} = the energy derived from enhanced van der Waals interactions in the complex. ΔG_{H} = the free energy attributed to the hydrophobic effect (0.125 kJ/mol per Å^2 of hydrocarbon surface removed from the solvent by complex formation).

* To whom correspondence should be addressed.

† Current address: Centre for Drug Design and Development, University of Queensland, Brisbane, Queensland 4072, Australia.

‡ Current address: Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545.

§ Current address: Pharmacokinetics Branch, ETD/HERL/USEPA, Research Triangle Park, NC 27711.

⊥ Current address: School of Chemistry, University of Leeds, Leeds LS2 9JT, U.K.

⊗ Abstract published in *Advance ACS Abstracts,* April 1, 1996.

(1) Böhm, H.-J. *J. Comput.-Aided Mol. Des.* **1992,** *6,* 61−78.

(2) Ho, C. M. W.; Marshall, G. R. In *Proceedings of the twenty-seventh annual Hawaii International Conference on system sciences*; Hunter, L., Ed.; IEEE Computer Society Press: Washington, DC, 1994; Vol. 5, pp 213−222.

(3) Pearlman, D. A.; Murko, M. A. *J. Comput. Chem.* **1993,** *14,* 1184−1193.

(4) Rotstein, S. H.; Murcko, M. A. *J. Med. Chem.* **1993,** *36,* 1700−1710.

(5) Goodford, P. J. *J. Am. Chem. Soc.* **1985,** *28,* 849−56.

(6) Miranker, A.; Karplus, M. *Proteins: Struct. Funct., Genet.* **1991,** *11,* 29−34.

(7) Lauri, G.; Bartlett, P. A. *J. Comput.-Aided Mol. Des.* **1994,** *8,* 51−66.

(8) Ho, C. M. W.; Marshall, G. R. *J. Comput.-Aided Mol. Des.* **1993,** *7,* 623−647.

(9) Ajay; Murcko, M. A. *J. Med. Chem.* **1995,** *38,* 4953−4967.

(10) Williams, D. H.; Cox, J. P. L.; Doig, A. J.; Gardner, M.; Gerhard, U.; Kaye, P. T.; Lal, A. R.; Nicholls, I. A.; Salter, C. J.; Mitchell, R. C. *J. Am. Chem. Soc.* **1991,** *113,* 7020−7030.

(11) Williams, D. H. *Aldrichim. Acta* **1991,** *24,* 71−80.

(12) Doig, A. J.; Williams, D. H. *J. Am. Chem. Soc.* **1992,** *114,* 338−343.

(13) Searle, M. S.; Williams, D. H. *J. Am. Chem. Soc.* **1992,** *114,* 10690−10697.

(14) Searle, M. S.; Williams, D. H.; Gerhard, U. *J. Am. Chem. Soc.* **1992,** *114,* 10697−10704.

(15) Searle, M. S.; Westwell, M. S.; Williams, D. H. *J. Chem. Soc., Perkin Trans. 2* **1995,** 141−151.

Searle and Williams[13] have examined the thermodynamics of sublimation of organic compounds without internal rotors, and shown that only 40−70% of theoretical entropy loss occurs on crystallization. This provides an estimate of the entropy loss to be expected on drug−ligand interaction. Pickett and Sternberg[16] have analyzed the side chain conformational entropy change on protein folding for each type of amino acid residue. As several of the components in the binding energy estimate are directly related to the degree of order of the system (entropy), simulations in solvent may be necessary to quantitate the degree by which the relative motions of the ligand and protein are quenched and the restriction on rotational degrees of freedom upon complexation. One might, therefore, consider thermodynamically rigorous approaches in which a ligand of known structure and affinity is mutated to the ligand of interest and the difference in binding free energy is calculated.[17−19] A good example of the use of simulations to calculate the relative free energy of binding is that of a transition state inhibitor to the enzyme thermolysin.[20] Such methods are appropriate when limited structural variation is under consideration as they have been shown accurate when the perturbation to the starting ligands is small. Unfortunately, such thermodynamic cycle perturbation approaches are computationally demanding and currently inappropriate for de novo design due to the diversity of the structures to be considered.

Quantitative structure−activity relationships (QSAR) represent another approach to affinity prediction. In these methods the thermodynamics of binding are not explicitly represented, but are embedded in physicochemical properties determined for each ligand which are then correlated to activity. Traditional QSAR methods have been developed for receptors whose 3D structure are unknown to analyze the database of ligands whose structures and binding affinities are known.[21] These are often based on the assumption that a correlation exists between the enthalpy of binding and the free energy of binding as the receptor is a constant and the congeneric series of ligands under consideration do not differ significantly in size, flexibility, etc. which would impact the entropy of binding. 3D QSAR methods, such as comparative molecular field analysis (CoMFA),[22−26] often use a grid-based approach to derive an active site model for the receptor in terms of energetic fields such as electrostatics, sterics, and hydrophobicity.[27] The difficulty with such methods lies within the limitations on the predictive models they generate and the requirements for their appropriate use. First, a considerable training set of ligands of diverse structure with known binding affinities must be present for each receptor. New therapeutic targets generally lack such

a set of diverse ligands with measured affinities. Second, the accuracy of prediction of ligands has been generally shown to be dependent upon the similarity of the ligand with those in the training set.[28] There is little confidence in the prediction of the binding affinity of a novel ligand which is truly unique with respect to the training set due to the extrapolation required from ligands present in the training set. In an extension to receptors of known structure, 3D QSAR has been used to derive robust predictive models for the binding affinity of inhibitors of HIV protease,[29−32] but one can argue that most of these compounds are of similar nature, being derived by substrate modification and retaining a peptidic backbone.

In an effort to overcome the limitations and approximations of QSAR approaches, a new class of scoring function with greater range of applicability has been developed. This type of function uses the receptor's 3D structure along with that of the ligand to predict affinity.[33−35] Such approaches, e.g., the scoring function for LUDI,[36] base their calculations upon an estimate of the binding free energy by approximating the contributions of hydrogen bonding, of entropy due to rotatable bonds in the ligand which are frozen upon binding, and of desolvation based on some sort of hydrophobic complementarity information. Böhm[36] analyzed 45 protein−ligand complexes (affinity range −9 to −76 kJ/mol) and found the following equation by multiple regression analysis:

$$\Delta G_{\text{binding}} \text{ (kJ/mol)} = 5.4\Delta G_{\text{o}} - 4.7\Delta G_{\text{hb}} - 8.3\Delta G_{\text{ionic}} - \\ 0.17\Delta G_{\text{lipo}} + 1.4\Delta G_{\text{rot}}$$

$$r^2 = 0.76, S = 7.9, q^2 = 0.696, s(\text{press}) = 9.3 \text{ kJ/mol}$$
$$(2.2 \text{ kcal/mol})$$

where ΔG_{o} is related to the reduction in rotational and translational entropy, ΔG_{hb} is the free energy associated with hydrogen bond formation, ΔG_{ionic} is the binding energy from ionic interactions, ΔG_{lipo} is the lipophilic interaction contribution, and ΔG_{rot} is the energy loss by freezing of internal degrees of freedom in the ligand.

Similarly, Krystek et al.[37] analyzed 19 protein−ligand complexes in an update of the Novotny approach to binding entropy[38] and produced the following relationship:

$$\Delta G_{\text{binding}} \text{ (kcal/mol)} = 11 - 0.025\Delta G_{\text{CSA}} - \Delta G_{\text{EL}} + 0.6T_{\text{sc}}$$

$$r^2 = 0.69, s = 4.0$$

where ΔG_{CSA} = the hydrophobic energy due to loss of surface

(16) Pickett, S. D.; Sternberg, M. J. *J. Mol. Biol.* **1993**, *231*, 825−39.

(17) Jorgensen, W. L.; Ravimohan, C. *J. Chem. Phys.* **1985**, *83*, 3050−3054.

(18) Straatsma, T. P.; McCammon, J. A. *Methods Enzymol.* **1991**, *202*, 497−511.

(19) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.

(20) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* **1987**, *235*, 574−6.

(21) Martin, Y. C. *Quantitative Drug Design: A Critical Introduction*; Marcel Dekker Inc.: New York, Basel, 1978; pp 425.

(22) Marshall, G. R.; Cramer, R. D., III. *Trends Pharmacol. Sci.* **1988**, *9*, 285−289.

(23) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(24) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. *J. Am. Chem. Soc.* **1993**, *115*, 5372−5384.

(25) *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 1−759.

(26) Green, S. M.; Marshall, G. R. *Trends Pharmacol. Sci.* **1995**, *16*, 285−291.

(27) Abraham, D. J.; Kellogg, G. E. In *3D QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 506−522.

(28) Folkers, G.; Merz, A.; Rognan, D. In *3D-QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 583−618.

(29) Waller, C. L.; Marshall, G. *J. Med. Chem.* **1993**, *36*, 2390−2403.

(30) Waller, C. L.; Oprea, T. I.; Giolitti, A.; Marshall, G. R. *J. Med. Chem.* **1993**, *36*, 4152−4160.

(31) Oprea, T. I.; Waller, C. L.; Marshall, G. R. *J. Med. Chem.* **1994**, *37*, 2206−2215.

(32) Oprea, T.; Waller, C.; Marshall, G. *Drug Des. Discovery* **1994**, *12*, 29−51.

(33) Horton, N.; Lewis, M. *Protein Sci.* **1992**, *1*, 169−181.

(34) Böhm, H.-J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(35) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; deSolms, S. J.; Gaffin, N.; Ghosh, A. K.; Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, W. M.; Tucker, T. J.; Wiggins, M.; Wiscount, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L.; Zugay, J. A. *J. Med. Chem.* **1995**, *38*, 305−317.

(36) Böhm, H.-J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(37) Krystek, S.; Stouch, T.; Novotny, J. *J. Mol. Biol.* **1993**, *234*, 661−679.

(38) Novotny, J.; Bruccoleri, R. E.; Saul, F. A. *Biochemistry* **1989**, *28*, 4735−4749.

area, ΔG_{EL} = the electrostatic binding energy, and T_{sc} = the change in torsional and rotational entropy.

Even these methods of approximating affinities have their limitations. First, they require the 3D structure of the receptor to make their predictions, although this is becoming less of an issue due to the advances mentioned earlier. Second, most of these methods employ calculations of limited accuracy due to computational constraints on estimation of the entropic contributions. We report a new hybrid methodology which focuses on the accuracy of the calculations through maximal use of 3D information of the ligand−receptor complex by combining a heuristic approach with parameters derived from molecular mechanics. (While this paper was in preparation, a method which uses parameters derived from molecular mechanics calculations on the ligand−receptor complex and linear regression analysis to develop a model of the affinity of ligands for the active site of a phospholipase was published by Ortiz et al.[39])

Theory

The processes governing molecular recognition and affinity consist of transfer of the ligand from the solvent to the more hydrophobic protein environment. The system undergoes changes in energy and entropy as a result of losing solute/solvent interactions, and the transfer of the ligand from solution to the active site. In most cases, the ligand has to displace water molecules which occupy the active site in its unliganded state. At the binding site, the ligand is conformationally immobilized, thereby decreasing its entropy and increasing its free energy. Simultaneously, optimized interactions with the functionality of the binding pocket increase the enthalpy of binding. The process can involve substantial conformational changes of both the ligand and the receptor. An absolute prerequisite for specific binding, however, is steric complementarity. Irrespective of the pattern of intermolecular forces surrounding the ligand, if a bulky group is present that does not fit within the three-dimensional active site available at the receptor, then effective binding is precluded due to the magnitude of the repulsive van der Waals interaction.

In summary, the success of a particular recognition event is dependent upon three major interactions that involve changes in both entropy and enthalpy. The first is the conformational changes of the ligand and the receptor upon complexation. The second is the structural and energetic complementary of the ligand and the receptor in the complex. The third is the thermodynamic aspects which describe the transfer of the ligand from solution to the binding site, i.e., the desolvation of ligand and receptor, and the loss of rotational and translational entropy of the ligand. In order to successfully predict the affinities of ligands for receptors, one must attempt to quantitate these events and scale their relative contributions to the free energy of binding.

Conformational Properties of the Ligand and Receptor. Changes in conformational entropy occur when the freely rotating side chains of the dissociated components are forced to adopt more rigid conformations on complex formation. Novotny et al.[38] have attempted to take entropy changes into consideration in free energy calculations of antibody−antigen complexes. The minimal estimate of conformational freedom lost assumes that each torsional degree of freedom has approximately three equivalent energetic states available, namely,

the trans and ± gauche. To estimate the total change in side chain conformational entropy (ΔS_{CF}), the atoms involved in the contact area of the complex were used to estimate the number, N, of side chain torsions fixed:

$$\Delta S_{CF} \text{ (kJ/mol)} = -R \ln(3^N) = -NR \ln 3 = -2.18N$$

Williams et al.[10] have estimated the entropy change due to freezing a free rotor to be 5 kJ/mol. Similarly, an indication of the conformational entropy lost upon binding[40] can be calculated by estimating the flexibility of the ligand. The flexibility index developed by Fisanick et al.[41] is a function of the shortest topological paths between all pairs of atoms in a structure, and takes into account the type of bonds and the extent of branching in the paths. An alternative measure of this is to simply count rotatable bonds.

In VALIDATE, the number of rotatable bonds in the ligand is counted by summing all nonterminal single bonds (methyl groups are assumed to be freely rotating). It has been suggested that the number of degrees of freedom in a nonaromatic ring system (aromatic rings are ignored) is on the order of $n - 6$,[42] where n represents the number of bonds in the ring. For this application, it was felt that $n - 4$ is more appropriate. The impetus for this is based upon the conformational analysis of five-membered systems such as proline. To count such a ring as having zero degrees of freedom did not seem appropriate due to puckering of the ring. Therefore, the rotatable bond (rb) count can be expressed as follows:

$$\text{no. of rb} = (\text{no. of nonterminal single bonds}) + \sum_i (n_i - 4)$$

$$n_i = \text{no. of bonds in ring } i$$

There is one exception to this method. In protein−protein systems, a large portion of the protein inhibitor is not bound at the active site and remains freely accessible to the solvent. Rotatable bonds in these areas, therefore, cannot be considered as being more "frozen" by complex formation. Hence, only the rotatable bonds at the active site interface are counted. We ignore the estimation of the energy associated with entropy loss in an explicit sense and simply use the number of torsions which are in the region of interaction as a parameter for the learning paradigm.

Transfer of Ligand from Solution. From a physicochemical viewpoint, the lipophilicity of a hydrophobic molecule is estimated by the energy needed to create a cavity in the aqueous solvent in which the solute can fit. Jackson and Sternberg[43] have developed a scoring function for docking proteins which uses molecular surface area changes in combination with electrostatic free energy and side chain conformational entropy changes to distinguish near-native from non-native dockings of six protein complexes. With hydrophobic molecules, when the molecule binds to the receptor, the energy of cavity creation is released, entropically favoring the recognition process. log P is used as a measure of molecular lipophilicity in water (with reference to octanol). We use the fragment-based H log P[44] method in Hint 1.1[45,46] to compute the ligand's partition

(38) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. *J. Med. Chem.* **1995**, *38*, 2681−2691.

(40) Murphy, K. P.; Xie, D.; Thompson, K. S.; Amzel, L. M.; Freire, E. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 63−67.
(41) Fisanick, W.; Cross, K. P.; Rusinko, A. *Tetrahedron Comput. Methodol.* **1990**, *3*, 635−652.
(42) Go, N.; Scheraga, H. A. *Macromolecules* **1970**, *3*, 178−187.
(43) Jackson, R. M.; Sternberg, M. J. *J. Mol. Biol.* **1995**, *250*, 258−75.
(44) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley and Sons: New York, 1979.
(45) Wireko, F. C.; Kellogg, G. E.; Abraham, D. J. *J. Med. Chem.* **1991**, *34*, 758−767.

coefficient. In essence, this value is used to gauge the ligand's preference for the active site of the receptor vs the aqueous solvent. Obviously, if it is energetically more favorable to remain in the solvent, few ligands will reach the active site. With the partition coefficient, a negative value indicates a preference for a polar (hydrophilic) environment and a positive value indicates a preference for a nonpolar (lipophilic) environment. When developing this method, we felt it was not sufficient to simply add the partition coefficient without using the additional knowledge of the receptor. For example, the HIV protease active site, though charged, is predominantly lipophilic in terms of total surface area. Thus, ligands with a positive partition coefficient are favored for the HIV protease binding site. However, in the case of the L-arabinose sugar binding protein, the active site is predominantly hydrophilic in total surface area. The sugar ligands which bind to this protein are hydrophilic (negative partition coefficients). In VALIDATE, we compute the amount of hydrophilic and lipophilic surface area as ratios to the total surface area of the receptor active site. The final value of the partition coefficient is then modified on the basis of this information in the following fashion:

$$PC = (RC)(H \log P)(PC)$$

where $(H \log P)(PC$ = the partition coefficient as computed by Hint 1.1 and RC = 1 if the receptor active site is predominantly lipophilic and -1 if the receptor active site is predominantly hydrophilic.

The determination of the lipophilic/hydrophilic preference of the receptor active site is slightly more involved than simply comparing surface areas. The calculation of lipophilic and hydrophilic surface areas in the active site is done using each ligand that binds with the particular receptor as a seed. The ligand is placed in the binding site and each receptor atom which has a solvent accessible surface that is within a distance R (where R is the mean radius of the solvent; water with a mean radius of 1.4 Å was used for the calculations in this paper) of the solvent accessible surface area of any ligand atom is considered part of the active site. The solvent accessible surface area of this atom is added to the appropriate surface area type. The definition of the surface area types should be made here. Any carbon which is covalently bonded to no more than one non-carbon is considered lipophilic. Any hydrogen connected to such a carbon is also considered lipophilic. All other atoms are considered hydrophilic. This is based upon the definition used in the Böhm scoring function.[34]

The determination of the sign of the coefficient RC is based on the sums of the lipophilic and hydrophilic surface areas of the active site. There are two considerations at this point. First, an active site that has slightly more hydrophilic than lipophilic surface area is still more lipophilic than a solvent like water. Second, the surface area calculations are highly dependent upon the size and shape of the ligands. The rest of the calculation keeps these considerations in mind and is, therefore, slightly biased to considering an active site as lipophilic. For the active site of a receptor to be considered hydrophilic and the value of RC to be set to -1, one of the following criteria must be true. If less than five ligands are available for a given active site, all calculations must yield that at least 55% of the total surface of the active site is hydrophilic. If five or more ligands are available, at least half of the calculations must yield that 55% or more of their total surface area is hydrophilic. The remaining calculations must yield that at least a majority of the total surface area is hydrophilic. Although there is no obvious precedent

for these criteria, they are derived from observations on the training set of 51 crystalline complexes processed. While a less pragmatic rationale is desirable, most systems we have examined thus far exhibit a substantially greater surface area of one type than the other, and no question arises as to the preference of the active site.

Protein ligands require a special treatment for this calculation as they did for the rotatable bond count. Since only a small portion of the protein ligand, the active site, is desolvated by binding to the receptor, only the $H \log P$ for this region is relevant. Therefore, this region is extracted from the protein and the calculation is done only on this part of the molecule.

Steric and Energetic Complementarity. An absolute prerequisite for specific binding is structural complementarity of the ligand and receptor at the site of interaction. Even though one could argue that hydrogen bonds and van der Waals (VDW) contacts contribute little to protein complexation[47] due to the competitive interactions with water in the unliganded state, there must be interaction between complementary desolvated surfaces, such that polar atoms are properly positioned to make hydrogen bonds. That is, electrostatic interactions contribute to the specificity of the complex formation; incorrect associations are forbidden by large unfavorable enthalpies due to poor packing and loss of hydrogen bonds made to water.

The nonbonded electrostatic interaction energy is calculated using the explicit sum of the Coulombic potentials:

$$\Phi = \frac{1}{4\pi\epsilon_0}\sum_i^L\sum_j^R\frac{q_iq_j}{r_{ij}}$$

The charges on the ligand and the receptor are those from the implementation of the Amber force field within the Macro-Model[48] program. Additions to this Amber charge set in MacroModel allow for appropriate charges for non-amino acid structures. The nonbonded steric interaction energy is computed from the explicit sum of the Lennard-Jones potentials:

$$E_{VDW} = \sum_i^L\sum_j^R\epsilon_{ij}(1/R_{12} - 2/R_6)$$

where

$$\epsilon_{ij} = (\epsilon_i\epsilon_j)^{1/2}$$

$$R_{12} = (r_{ij}/(R_i + R_j))^{12}$$

$$R_6 = (r_{ij}/(R_i + R_j))^6$$

r_{ij} = the distance between atom center i and atom center j, R_i, ϵ_i = the VDW radius and epsilon value of atom i, and R_j, ϵ_j = the VDW radius and epsilon value of atom j. The parameters required were derived from the Amber force field within the MacroModel program.

Steric Fit. It has been demonstrated that individual amino acid residues in proteins are closely packed[49] with individual amino acid residues occupying the same volume as they do in crystals of the amino acids themselves.[47,50] This has resulted in the interfaces between residues in protein hydrophobic cores

(46) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 545−552.

(47) Chothia, C.; Janin, J. *Nature* **1975**, *256*, 705−708.
(48) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. *J. Comput. Chem.* **1990**, *11*, 440−467.
(49) Richards, F. M. *J. Mol. Biol.* **1974**, *82*, 1−14.
(50) Chothia, C. H. *Nature* **1975**, *254*, 304−308.

being described as more "solid"-like.[51] Surface complementarity is, therefore, an important component of recognition.[52] Thus, we have developed a function called "steric fit" which describes the close packing interactions between the ligand and receptor. The steric fit is computed by summing the number of "good contacts" for each atom of the ligand which is contained in the active site. Antibody−steroid complexes and chymotrypsin-binding proteins actually contain a considerable percentage of their atoms outside of the active site while HIV protease binding inhibitors are largely surrounded by the receptor. VALIDATE defines a good contact as an instance where the VDW surface of a ligand atom is within a modifiable parameter, ϵ, of the VDW surface of a receptor atom:

$$\text{SF} = (\sum_{i,j} C_{ij})/N$$

where

$$C_{ij} = 1, d_{ij} \leq |r_i + r_j \pm \epsilon|$$
$$0, d_{ij} > |r_i + r_j \pm \epsilon|$$

$N =$ the number of ligand atoms contained in the active site, r_i, $r_j =$ the VDW radii of atoms i and j, and $d_{ij} =$ the distance from atom center i to atom center j.

We have examined ϵ values ranging from 0.1 to 0.3 Å. In actuality, the range is not necessarily symmetric as the equation would indicate. For obvious reasons, when $\epsilon > 0.1$ Å, we do not allow an overlap ($-\epsilon$) of greater than 0.1 Å even though the allowed gap ($+\epsilon$) could be defined to be as large as 0.3 Å. The results reported in this paper were computed using $\epsilon = 0.3$ Å. It is important to note the scaling of this value by the number of ligand atoms within the active site region.

Surface Areas. Previously, electrostatic and hydrophobic properties have been mapped onto contoured surface areas as an aid to drug design[53] as the two interacting surfaces must be sterically and electrostatically complementary. In this regard, VALIDATE computes four components to surface complementarity. These are lipophilic complementarity (nonpolar/nonpolar), hydrophilic complementarity (polar/polar, opposite charge), lipophilic/hydrophilic noncomplementarity (polar/nonpolar), and hydrophilic noncomplementarity (polar/polar, like charge). VALIDATE uses 256 evenly distributed data points, obtained from the SASA program,[54] which are placed on the VDW surface of each receptor atom whose VDW surface is within 5 Å of the atom center of any ligand atom. If a point on this surface is within a mean solvent radius (1.4 Å for water) of the VDW surface of a ligand atom, it is considered a contact point. Its type is based upon the determination of the polar/nonpolar nature of both atoms and the criteria discussed above. At this point, we split into two different types of calculations of surface area. The first is an absolute surface area between the ligand and receptor similar to the method used by Böhm.[34] For each point on the receptor surface, a record of each type of contact is kept. However, a particular type of contact, e.g., lipophilic complementarity, is counted only once even if that point is within the distance limit described above of more than one ligand atom's VDW surface. The points of each type are then summed for each atom in the receptor. The total surface on each atom for each type of contact is computed by dividing the number of contact points of that type by 256 (the total number

of points possible) and then multiplying by the total surface area of the atom. The functional form is

$$\text{CSA} = \sum_{i}^{R} (4\pi r_i^2 (\text{CP}_i))/256$$

where $\text{CP}_i =$ the number of contact points on atom i and $r_i =$ the VDW radius of atom i. The second method is somewhat similar to the approach in Hint 1.1[45,46]—a pairwise sum. If a single point on the surface of a receptor atom is within the described distance of n atoms of the ligand of a given type, then that point would record a sum of n, rather than 1 as previously described. In this paper, the results of both types of calculations are included. While we expected one or the other procedure to be preferred, using both sets of data in the model improved not only the model itself but its predictive abilities on all test sets by a significant amount (on the order of 0.1 unit). We assume this reflects surface complementarity in some nonlinear manner related to exclusion of solvent.

When these fields were initially added to the model, there was a good correlation among ligands associated with a given receptor; however, there appeared to be a scaling problem between receptor types. To address this problem, it was decided to scale the values computed for surface areas on the basis of the ratio of the total surface area of each receptor's active site to that of HIV protease. HIV protease was chosen as a convenient normalization factor as there were more HIV inhibitors in the training set than any other type. To do the scaling, the largest total surface area computed for a given receptor's active site for any of its known ligands is used to divide the largest total surface area computed for HIV protease. The scaling factor is used to multiply each lipophilic and hydrophilic contact surface area. An attempt to use the average total surface area of the active site was explored; however, the results were not as good as those using the largest area for normalization. Due to the method of scaling, it should make no difference in the overall results which receptor was used as a reference since all numbers are relative.

Ligand Strain Energy. Ligand strain energy may be viewed as the amount of energy required for the ligand to adopt the receptor-bound conformation. We define the strain energy pragmatically by

$$\text{IE} = |E_{\text{bs}} - E_{\text{solv}}|$$

where $E_{\text{bs}} =$ the energy of the ligand's receptor-bound conformation and $E_{\text{solv}} =$ the energy of the ligand in solvent at its nearest local minimum. This is calculated by comparison of the conformational enthalpy of the receptor-bound conformation of the ligand to the nearest local minimum of the unbound ligand using the GB/SA[55] solvation model with the Amber all-atom force field implementation in MacroModel.

At this point, an overall summary of the model should be given. The major properties computed are the electrostatic energy, steric interaction energy, steric fit, $H \log P$ partition coefficient, rotatable bond count, ligand strain energy, lipophilic contact surface area, hydrophilic contact surface area, polar/nonpolar contact surface area, and bad hydrophilic contact surface area. A note should be made at this point about hydrogen bonds which are not explicitly considered as they are in the LUDI scoring function.[34] Initially, a hydrogen bond factor was computed using the method described by Böhm.[34] However, all attempts to incorporate this parameter into the current

(51) Bello, J. *Int. J. Peptide Protein Res.* **1971**, *12*, 38−41.
(52) Lawrence, M.; Colman, P. *J. Mol. Biol.* **1993**, *234*, 946−950.
(53) Bohacek, R. S.; McMartin, C. *J. Med. Chem.* **1992**, *35*, 1671−1684.
(54) Le Grand, S. M.; Merz, K. M., Jr. *J. Comput. Chem.* **1993**, *14*, 349−352.

(55) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

model reduced its accuracy. One should not conclude, however, that hydrogen bonds are not important to receptor−ligand complexation. Hydrogen bonds make a significant contribution to the interaction energy; however, they must be adequately represented by other parameters used in the model including the electrostatic interaction energy and the complementary polar surface area.

Computational Methods

The calculation of properties for validation is a three-step process. First, the crystal structure of the complex is processed into a usable form. Second, the complexes need to be minimized to alleviate steric contacts which usually arise due to the random way in which hydrogens are added to heavy atoms. Finally, the structures are processed through the VALIDATE program to compute the requisite parameters. When explained in detail, the process sounds quite involved and time consuming. However, all steps, except for the initial extraction of ligand from the complex, have been completely automated through software. Extracting the receptor and ligand from the cocrystal is, in fact, the most time-consuming effort in processing. Most of the structures of complexes are from Brookhaven;[56] however, some are from other sources. The Brookhaven files are initially read into Sybyl[57] where the ligand is extracted from the receptor−ligand complex.

Hydrogens are added to the receptor with the biopolymer dictionary in Sybyl. Since most of the ligands are not peptides, the hydrogens are added through valence-filling algorithms. Waters and any ions in the crystal structures of the complex were left with the receptor. The ligand and the receptor are then written out as separate files in Sybyl's mol2 format. Due to the way hydrogens have been added to the receptor, ligand, and water molecules, it was necessary to minimize the complexes before computing the VALIDATE fields. It was our intent to simply relieve steric contacts created from this addition, not to have large conformational changes occur during minimization. The criteria used to set up the minimization helps to ensure that this was the case. The minimizer used was the Batchmin minimizer distributed with MacroModel.[48] Several force field implementations are available with this minimizer, and we used the Amber all-atom implementation. This force field was selected because it was generally well parameterized for most of our complexes. The setup criteria for the minimization basically involved selecting a core of atoms around the active site for minimization and a shell of atoms surrounding that to be used for long-range electrostatics interactions during the minimization. Atoms in the ligand and the selected core of the receptor underwent the full minimization process. Atoms in the shell surrounding this were fixed in place, but charges were placed on them by the minimizer to be used for nonbonded electrostatic interactions with atoms in the core and ligand. Atoms beyond this shell were ignored. In the software that generates the files for minimization, the sizes of the core and the shell are specified in angstroms. For all complexes in this paper, the values of 8 and 10 Å were used for the core and the shell, respectively. The first distance was used as a cutoff based on the atoms of the ligand and is not a spherical radius. Any receptor atom within 8 Å of any ligand atom was added to the core. Any atom within 10 Å of these selected core atoms was added to the fixed shell. It should be noted that waters and ions are included when selecting these atoms. An illustration of the core and shell can be seen in Figure 1. The minimizations were run with a gradient convergence of 0.1 and an iteration limit of 5000 for all small molecule ligands (15 000 for the protein inhibitors). A solvation model was not used; however, in many cases numerous waters from the crystal structures were included in the core and the shell. Minimizations were run on SGI R4000 Challenge M machines with the mips2 compiled version of the batchmin program from MacroModel 4.5. Average run times were on the order of 30−45 min (protein inhibitors took considerably longer).

Once a minimization was complete, the ligand and the receptor were extracted into separate Sybyl mol2 format files. The VALIDATE program was run with these two files as input. The electrostatic
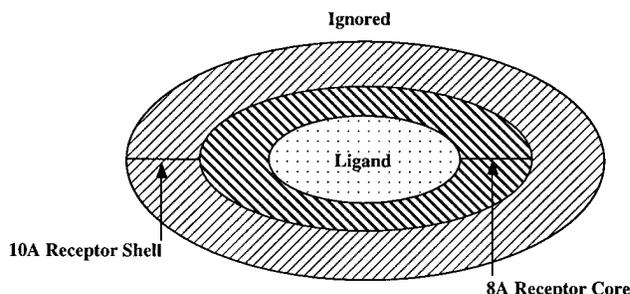

Figure 1. Illustration of receptor atom selection for minimization based on using the ligand as a seed.

interaction energy, steric interaction energy, steric fit, rotatable bond count, and all lipophilic/hydrophilic surface area calculations are done internally. VALIDATE calls the Hint program to compute the H log P and extracts the result from the Hint output file. All calculations for this paper have been done using Hint 1.1.[58] A secondary program calls batchmin to minimize the extracted ligand using the Amber all-atom force field implementation with the GB/SA solvation model. The starting energy and final energy are extracted from the batchmin output file, and the ligand strain energy is calculated. The run times for calculation of the properties computed internally in VALIDATE are generally on the order of 10−30 s. The H log P calculation averages about 10 s for a typical ligand. The ligand strain energy calculation varies from 1 to 10 min with an average of 5 min per ligand (protein ligands took considerably longer).

The fields which are computed by VALIDATE were placed into a molecular spreadsheet in Sybyl. The first column of the spreadsheet contained the binding affinity which is used as the dependent column of a PLS analysis[59,60] as implemented in Sybyl. The analysis produced a coefficient for each data field that was used for the prediction of new compounds. For a basis of comparison, the analysis was also performed using the SONNIC neural network program[61] with the same parameters as input. The results of both methods are presented in this paper.

Results

The training set consisted of the parameters calculated for 51 receptor−ligand cocrystalline complexes. The training set included complexes of HIV protease, thermolysin, endothiapepsin, β-trypsin, and subtilisn-Novo inhibitors, antibody (DB3)-bound steroids, and L-arabinose binding protein-bound sugars as listed in Table 1. Complexes were chosen to produce a set which was diverse in the areas of interaction size, type, activity, and physicochemical features. Most of the complexes were selected from a list compiled by Keske and Dixon[62] which contains cocrystal complexes, available in the Brookhaven database, along with their published binding affinities. Beyond attempting to achieve physicochemical diversity, the only criteria for elimination were systems which were too large to be computationally convenient and those which required additional parameterization for the force field. The diversity of the set is illustrated in the different types of receptors included. HIV protease and endothiapepsin complexes represent small molecular inhibitors bound to aspartyl proteases, thermolysin complexes are a metalloprotease with bound small molecular inhibitors, β-trypsin and subtilisin-Novo complexes have protein inhibitors bound in the complexes used, DB3 is a steroid-binding antibody, and L-arabinose binding protein complexes have a

(56) Abola, E. E.; Bernstein, F. C.; Koetzle, T. F. In *The Role of Data in Scientific Progress*; Glaeser, P. S., Ed.; Elsevier: New York, 1985.

(57) SYBYL, Tripos Associates, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

(58) HINT, eduSoft, LC, P.O. Box 1811, Ashland, VA 23005.

(59) Wold, S.; Albano, C.; Dunn, W. J., III; Esbensen, K.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjostrom, M. *Analysis* **1984**, *12*, 477−85.

(60) Wold, S.; Johansson, E.; Cocchi, M. In *3D QSAR in Drug Design*; H. Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 523−550.

(61) Broughton, H.; Green, S.; Rzepa, H. *J. Chem. Soc., Perkin Trans. 2* **1995**, 431−435.

(62) Dixon, J. S.; Keske, J. *New Methods Drug Res.* in press.

Table 1. Receptor−Ligand Complex List for the Base Model[a]

protein−inhibitor	Brookhaven	affinity (−log K_i)	protein−inhibitor	Brookhaven	affinity (−log K_i)
HIV−AG1001	N/A	4.62*	endothiapepsin−PD 125754	1eed	4.90
HIV−AG1002	N/A	4.91*	endothiapepsin−L-364,099	2er0	6.40
HIV−AG1004	N/A	5.28*	endothiapepsin−H 256	2er6	7.20
HIV−RochIV	N/A	7.52*	endothiapepsin−H 261	2er7	9.00
HIV−MVT101	4hvp	6.12	endothiapepsin−L-363,564	2er9	7.40
HIV−SC52964	N/A	6.65*	endothiapepsin−CP 71,362	3er3	7.10
HIV−JG365	7hvp	9.60	endothiapepsin−PD 125967	4er1	6.60
HIV−acetylpepstatin	5hvp	5.60	endothiapepsin−H 142	4er4	6.80
HIV−GR116624X	N/A	7.05*	endothiapepsin−CP 69,799	5er2	6.60
HIV−U75975	N/A	8.65*	L-arabinose bind prot−L-arabinose	1abe	6.50
HIV−L-689,502	N/A	8.95*	L-arabinose bind prot−D-fucose	1abf	5.20
HIV−A74704	9hvp	8.50	L-arabinose bind prot P254G−D-fucose	1abp	5.80
HIV−A77003	N/A	9.70	L-arabinose bind prot P254G−L-arabinose	1bap	6.90
HIV−hydroxyethylene	1aaq	5.50	L-arabinose bind prot P254G−D-galactose	9abp	8.00
HIV−L-700,417	4phv	9.15	L-arabinose bind prot M108L−L-arabinose	9abp	7.00
thermolysin−phosphoramidon	1tlp	7.55	L-arabinose bind prot M108L−D-fucose	7abp	5.40
thermolysin−N-(1-carboxy-3-phenyl)-L-LeuTrp	1tmn	7.47	L-arabinose bind prot M108L−D-galactose	8abp	6.60
thermolysin−N-phosphoryl-L-leucinamide	2tmn	4.10	β-trypsin−BPTI	1tpa	14.00
thermolysin−ValTryp	3tmn	5.90	β-trypsin−PTI	2ptc	13.30
thermolysin−Leu-NHOH	4tln	3.72	DB3−11a-hemisuccinate	N/A	9.44
thermolysin−ZFPLA	4tmn	10.19	DB3−5a,3b-hemissuccinate progesterone	N/A	8.70
thermolysin−ZGp(NH)LL	5tmn	8.04	DB3−AEtiocholanolonE	N/A	7.62
thermolysin−ZGp(O)LL	6tmn	5.05	DB3−progesterone	N/A	9.00
thermolysin−CH_2CO-Leu-OCH_3	7tln	2.47	subtilisin-Novo−Eglin c L45R	1sbn	10.30
endothiapepsin−PD 125754	1eed	4.90	subtilisin-Novo−CI-2	2sni	11.00
endothiapepsin−L-364,099	2er0	6.40	subtilisin-Novo−SSI M73K	3sic	10.20
endothiapepsin−H 256	2er6	7.20	subtilisin-Novo−SSI M70G M73K	5sic	10.20

[a] An asterisk indicates the activity is derived from IC_{50} values by use of the Cheng and Prusoff equation.

predominantly hydrophilic active site with bound sugars. The ligands in this set ranged from 24 atoms (Leu-NHOH) to 1512 atoms (SSI M70G M73K), and the activity range extended from −log K_i = 2.47 to −log K_i = 14.0.

The statistical (PLS) and neural network (SONNIC) analyses produced similar results with respect to the correlation values. The PLS analysis yielded $r^2 = 0.849$ with $s = 1.006$ log units. A leave-one-out cross-validation of the set produced $q^2 = 0.776$ with $s(press) = 1.139$ with six components. Six components had been chosen on the basis of previous experience with these parameters as being the most robust. For this particular training set, $s(press)$ was equivalent for the four-component model. The SONNIC analysis trained to $r^2 = 0.81$. Cross-validation yielded $q^2 = 0.765$ at this level of training and a rms error of 1.08. The full results of the PLS and SONNIC analysis are listed in Table 2a. The graphs of actual vs predicted binding affinity in −log K_i for the fitted and cross-validated PLS analyses are shown in Figures 2 and 3. The field coefficients and their contribution percentage to the overall model are listed in Table 2b. The r^2, q^2, s and $s(press)$ are those given in the PLS analysis done within Sybyl and defined as follows:

$$q^2 = 1 - \left(\frac{\sum_N (\text{pred}_i - \text{actual}_i)^2}{\sum_N (\text{actual}_i - \bar{X})^2} \right)$$

$$s = \left[\frac{\sum_N (\text{calc}_i - \text{actual}_i)}{(N - c - 1)} \right]^{1/2}$$

$$s(press) = \left[\frac{\sum_N (\text{pred}_i - \text{actual}_i)}{(N - c - 1)} \right]^{1/2}$$

where N is the number of molecules and c is the number of
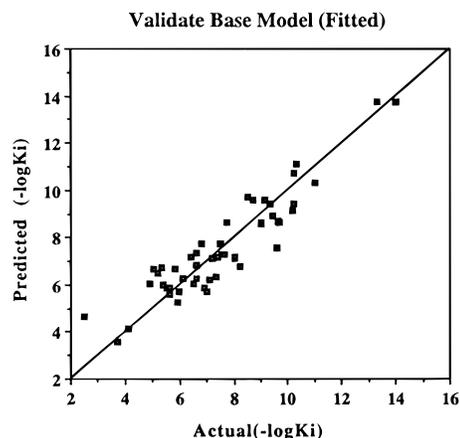

Validate Base Model (Fitted)

Figure 2. Fitted PLS analysis of 51 complexes in the training set. $r^2 = 0.85$ and the standard error $(s) = 1.01$.

PLS components (in the case of multiple regression analyses, c = the number of variables).

The true measure of any model rests in its ability to predict new compounds. For a generic model useful for de novo design, the model should obviously predict unique ligands for the receptors that exist in the training set, but more importantly, unique ligand−receptor complexes. Three separate test sets were compiled to test the derived model's predictability. The first test set consisted of 14 inhibitors which were obtained from crystalline receptor−ligand complexes. Neither the ligands nor the specific receptors in this test set were included in the training set. Included were two DHFR, two penicillipepsin, three carboxypeptidase, two α-thrombin, two trypsinogen, and three DNA complexes. Prediction of the binding affinities for this test set of novel complexes (mean binding affinity 6.90 versus 7.495 for the training set) was very good with a predictive $r^2 = 0.806$, an absolute average error of 0.697 (approximately 0.95 kcal/mol at 25 °C), and an rms error of 0.899 log unit (approximately 1.2 kcal/mol at 25 °C). On this test set,

Table 2

(a) PLS Analysis Results for the Base Set (Mean Binding Affinity 7.60)
Fitted Analysis

PLS	
standard error of estimate	1.01
r^2	0.85
F values ($n1 = 12$, $n2 = 38$)	17.81
prob of $r^2 = 0$ ($n1 = 12$, $n2 = 38$)	0.000
SONNIC	
trained r^2	0.81

Cross-Validation
PLS
Standard Error of Predictions (Cross-Validated)[a]

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6
$-\log K_i$	1.722	1.315	1.194	1.133	1.138	1.139

	comp 7	comp 8	comp 9	comp 10	comp 11	comp 12
$-\log K_i$	1.204	1.231	1.246	1.291	1.317	1.336

q^2	0.78
rms error	1.14
SONNIC	
q^2	0.77
absolute average error	0.89
rms error	1.08

(b) Coefficients and Contributions of Fields to the Base Model from PLS

field[b]	norm coeff	fraction
(1) electrostatic interaction energy	0.110	0.027
(2) steric interaction energy	0.440	0.107
(3) steric fit	0.034	0.008
(4) $H \log P$	0.339	0.082
(5) rotatable bonds	0.497	0.121
(6) induction enthalpy	0.479	0.117
(7) M1 lipophilic contact surface area	0.395	0.096
(8) M1 hydrophilic contact surface area (UnComp)	0.393	0.096
(9) M2 lipophilic contact surface area	0.759	0.185
(10) M2 hydrophilic contact surface area	0.334	0.081
(11) M2 lipophilic/hydrophilic contact surface area	0.223	0.054
(12) M2 hydrophilic contact surface area (UnComp)	0.106	0.026

[a] The optimum number of components is six. [b] M1 = method 1 and M2 = method 2 as described in the Theory section.
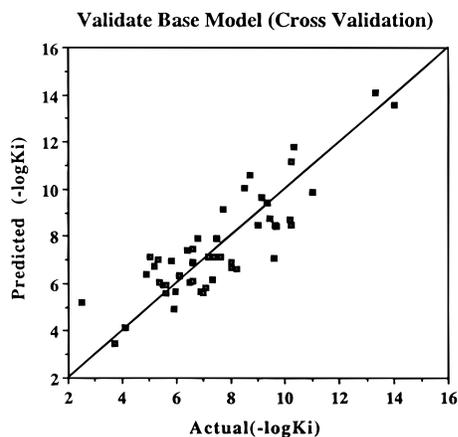
Validate Base Model (Cross Validation)

Figure 3. Cross-validated PLS analysis of 51 complexes in the training set. $q^2 = 0.78$ and the standard error (s(press)) $= 1.14$ with six components.

SONNIC scored a maximal predictive $r^2 = 0.85$ and an absolute average error of 0.96 log unit for training to a value of $r^2 = 0.84$ for the training set (Figure 7). The full results of the analysis can be seen in Table 3 and Figure 4. The predictive r^2 for this and all other test sets was computed analogous to the definition of r^2 by Cramer et al.[23] to discern the improvement of the prediction over the obvious prediction of the mean of the training set for each member of the test set which would

Table 3

(a) Listing of Compounds and Binding Affinities of Test Set 1[a]

protein−inhibitor	actual affinity ($-\log K_i$)	predicted affinity ($-\log K_i$)
DHFR−folate	7.4	7.29
DHFR−methetrexate	8.3	6.40
penicillipepsin−IvaVVLySta-OEt	9.4	7.71
penicillipepsin−Iva VVSta-OEt	7.7	8.04
carboxypeptidase−L-benzylsuccinate	6.3	5.92
carboxypeptidase−GlyTyr	4.0	5.14
carboxypeptidase−ZAGp(0)F	9.1	9.39
α-thrombin−MD 805	7.4	7.00
α-thrombin−NAPAP	8.2	8.54
trypsinogen−IleVal	4.3	4.35
trypsinogen−ValVal	2.9	3.39
DNA−daunomycin	6.5	6.10
DNA−netropsin	8.8	9.59
DNA−4,6-diamidino-2-phenylindole	6.3	5.04

(b) Results of Predictions of Test Set 1 (Mean Binding Affinity 6.900)

PLS	
predictive r^2	0.81
absolute average error	0.70
rms error	0.90
SONNIC	
predictive r^2	0.69
absolute average error	0.96
rms error	1.13

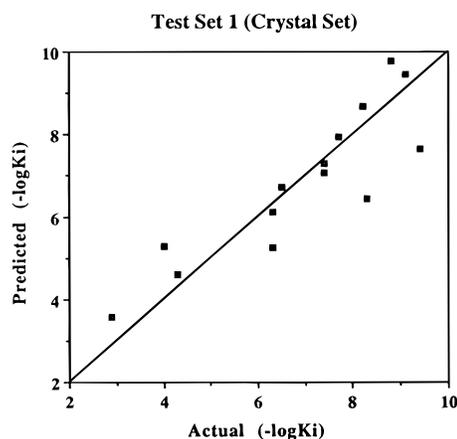[a] Predictions are from PLS analysis.

Test Set 1 (Crystal Set)

Figure 4. Prediction of affinities of 14 crystalline complexes using coefficients from fitted PLS analysis of the training set. The predictive $r^2 = 0.81$ and the absolute average error is 0.70.

give a predictive $r^2 = 0$ by the following formula:

$$\text{predictive } r^2 = (\text{SD} - s(\text{press}))/\text{SD}$$

where

$$\text{SD} = \sum_N (\text{actual}_i - \bar{X})^2$$

$$s(\text{press}) = \sum_N (\text{actual}_i - \text{pred}_i)^2$$

and $\bar{X}$ = the mean value of the binding affinity of the training set. The absolute average error is computed as

$$\text{absolute average error} = \frac{\sum_N |\text{pred}_i - \text{actual}_i|}{N}$$

Table 4

(a) Listing of HIV Protease Inhibitors and Binding Affinities in Test Set 2[a]

HIV inhibitor	actual affinity ($-\log K_i$)	predicted affinity ($-\log K_i$)
Z·Phe[CH(OH)CH$_2$N]Pro	4.84	6.55
QC·Asn·Peh[CH(OH)CH$_2$N]DIQ·NH^{t}Bu	9.05	8.24
Z·Phe[C(O)C(O)N]Pro·NH^{t}Bu	5.87	6.42
Z·Asn·Phe[CH(OH)C(O)N]Pro·NH^{t}Bu	5.16	6.06
NoA·His·Leu[CH(OH)CH(OH)]Val·Ile·Amp	7.60	7.23
m3	5.67	6.18
m4a	6.55	6.71
m7	6.23	6.22
m8a	7.10	6.47
m10a	7.19	6.98
m10b	7.95	8.14
m11b	8.17	7.64
m12	4.84	6.95

(b) Results of Predictions of Test Set 2 (Mean Binding Affinity 7.34)

	PLS
predictive r^2	0.57
absolute average error	0.73
rms error	0.87
	SONNIC
predictive r^2	0.53
absolute average error	0.72
rms error	0.90

[a] Predicted values are from the PLS analysis. K_i values are derived from published IC$_{50}$ values through use of the Cheng and Prusoff equation.

and the rms or root-mean-square error is computed as

$$\text{rms error} = \left[\frac{\sum(\text{pred} - \text{actual})^2}{N} \right]^{1/2}$$

The second test set consisted of 13 HIV protease inhibitors whose initial conformation and alignment were derived from the CoMFA analysis done by Waller et al.[30] The selection of the inhibitors was based upon maintaining a good range of activity (mean binding affinity 6.983 versus 7.495 for the training set) as well as utilizing several inhibitors from the published test set. A complete list is given in Table 4. All 13 inhibitors were minimized with the structure of HIV-1 protease extracted from the Brookhaven entry 9hvp. The PLS predictive r^2 was 0.568 with an absolute average error of 0.726 log unit and an rms error of 0.866 log unit (approximately 1.17 kcal/mol at 25 °C). The results of this analysis are shown in Figure 5. The SONNIC analysis yielded similar results with a maximal predictive $r^2 = 0.529$ and an absolute average error of 0.718 log unit at a training $r^2 = 0.83$ (Figure 7). The predictive r^2 is considerably lower than that of the first test set; however, this is due to the smaller range and values similar to the mean activity of the training set. The absolute average error is almost identical to that of the first test set.

The third and final test set was composed of the 11 thermolysin inhibitors used as a test set in the CoMFA analysis done by DePreist et al.[24] and Waller and Marshall.[29] The alignment of the inhibitors with respect to one another, as well as the initial conformation, is identical to that used by Waller and Marshall.[29] At this point, it should be noted that the Waller and Marshall alignment was a field-fit alignment. The set as a whole was then aligned to ZFPLA from the cocrystal complex. Several of the inhibitors in the test set contained a phosphorus group similar to ZFPLA, and this was used as the reference point for alignment. The structures were then minimized with the enzyme structure from the ZFPLA cocrystal. The PLS analysis yielded a predictive $r^2 = 0.715$ with an absolute average
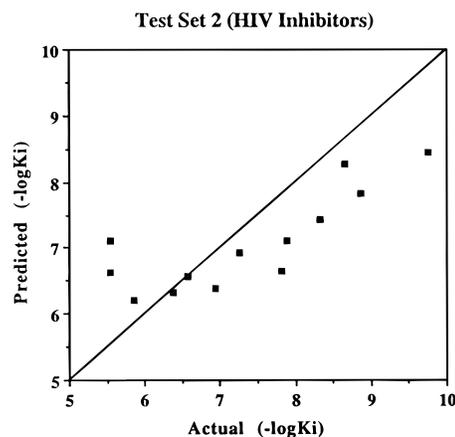

Figure 5. Predictions of affinities for 13 HIV protease inhibitors fit to the active site using coefficients from fitted PLS analysis of the training set. The predictive $r^2 = 0.57$ and the absolute average error is 0.73.

error of 1.48 log units. The high predictive r^2 reflects the large difference between the mean (7.495) of the training set and the mean (4.645) of test set 3. In other words, while the average errors for the thermolysin complexes were considerably larger than for the two other test sets, the model correctly predicted the overall lower affinity of these complexes compared with the average affinity of the training set. SONNIC performed approximately the same with a maximal predictive $r^2 = 0.70$ and an absolute average error of 1.50 log units at a training $r^2 = 0.87$ (Figure 7). The full analysis is listed in Table 5a. As can be seen in Figure 6, six of the inhibitors were predicted fairly well while the remaining five were not. The initial alignments discussed above produced several bad steric contacts, and starting energies were on the order of 10^7 kJ/mol. The results from this set will be discussed further in the next section.

Discussion

We have presented a new method for the prediction of binding affinity based on a hybrid model combining energetic considerations from molecular mechanics and calculated molecular properties related to desolvation and entropy loss on binding. A model for binding affinity was derived using 51 crystal structure complexes. A PLS analysis of this set provided coefficients which correlate the physicochemical properties with binding affinity. This mathematical model allows the prediction of the activities of new ligands at a given receptor, or novel ligand−receptor complexes, once the relevant parameters have been calculated. The results of the cross-validation of the training set as well as the predictions of the test sets are very encouraging. Even with the relatively poor performance on 5 of the 11 thermolysin inhibitors in test set 3, the absolute average error was only 1.45 log units. The performance on the first two test sets was very good with absolute average errors well under 1 log unit. While these results indicate that this technique is well suited to determine which drug design candidates are the best candidates for synthesis, it does not mean that further improvement is not possible through more judicious parameter selection and binding mode determination.

The similarity between the results of the PLS and neural network analyses is a partial indication that the model is consistent. However, the fields that were used in each analysis are slightly different. In order to achieve optimal results with PLS, the hydrophilic complementary contact surface area and the lipophilic/hydrophilic contact surface area columns using the first method of calculation were dropped in the PLS analyses

Table 5

(a) Listing of Thermolysin Inhibitors and Binding Affinities in Test Set 3[a]

thermolysin inhibitor	actual affinity ($-\log K_i$)	predicted affinity ($-\log K_i$)	thermolysin inhibitor	actual affinity ($-\log K_i$)	predicted affinity ($-\log K_i$)
PPPhe	7.20	7.23	ZGPOLF	4.30	7.17
PLFOH	3.50	4.06	ZGPOLG	3.64	6.04
ZFGNH2	6.57	6.36	ZGPOLNH2	3.18	5.18
ZGPLF	7.72	5.23	ZLGNH2	2.51	3.21
ZGPLG	2.70	5.69	ZYGNH2	3.66	2.75
ZGPLNH2	6.12	5.30			

(b) Results of Predictions of Test Set 3 (Mean Binding Affinity 4.65)

	PLS
predictive r^2	0.72
absolute average error	1.48
rms error	1.85

	SONNIC
predictive r^2	0.70
absolute average error	1.50
rms error	1.87

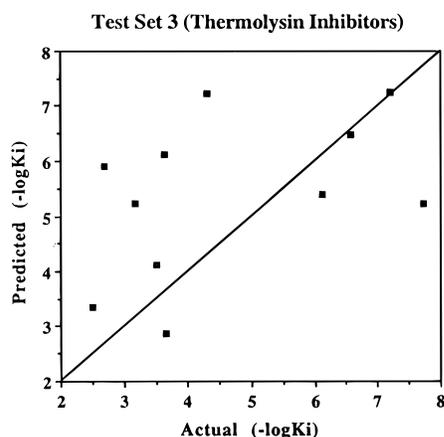[a] Predicted values are from the PLS analysis.

Figure 6. Prediction of affinities for 11 thermolysin inhibitors using coefficients from fitted PLS analysis of the training set. The predictive $r^2 = 0.72$ and the absolute average error is 1.48.

Figure 7. Predictive r^2 for the training set (leave-one-out) and three test sets (14 diverse complexes, 13 HIV protease inhibitors, and 11 thermolysin complexes) as a function of training using the cross-validated fit between the model and training set (trained r^2) as the objective function.

presented, leaving six parameters for surface interaction. These two columns apparently contain no new information for the model; however, the PLS implementation in Sybyl attempts to fit them anyway. It was not necessary to remove these two columns for the neural network analysis. SONNIC is capable of determining when a column is not contributing to the model and effectively ignores them; therefore, all columns were used as input. This is the one major advantage of using SONNIC over PLS. One difficulty in the use of neural networks is overtraining. In order to control for this problem, the cross-validated predictive r^2 of the training set is calculated and used to determine the maximal r^2 from training. In other words, the neural network model is trained while monitoring the predictive r^2, and training is terminated after the maximal predictive r^2 is found. Figure 7 indicates that the model is optimally trained at the trained $r^2 = 0.81$. The predictability of the model for the three test sets continues to rise with further training, reaches a maximum for the HIV protease inhibitor test set, for example, at trained $r^2 = 0.83$, and then degrades with further training. The predictive ability for the other two data sets actually increases to a maximum when the trained $r^2 = 0.85$ (diverse set) or 0.87 (thermolysin). The practical question, of course, is when to stop training. We would suggest that model derivation use the cross-validated r^2 of the training set to monitor overtraining. Our experience suggests that this maximum occurs in a range centered around a trained $r^2 = 0.85$ for the types of
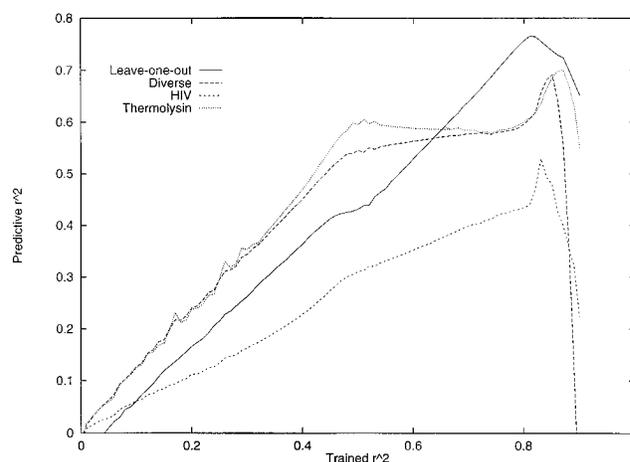
data sets we have examined. The discrepancy seen above in predictive results due to variations in the completeness of training suggests that the PLS approach is more robust, at least, at this stage of development of the two alternative approaches.

One of the major surprises of the model was the calculated role of the electrostatic interaction energy. The contribution of electrostatics to the model was only 2.9%. It is difficult to believe that the electronics of interaction contribute less than 3% to the binding process. In fact, this was anticipated to be one of the most significant physical properties of binding. Since most of the structures are derived from crystal structures, we can only assume either that our representation of charges is inadequate or that much of the relevant information from electrostatic interactions is found in other parameters such as complementary polar surface area. Some of the complexes contain bound zinc ions, and the force field parameterization we used is deficient for this metal, resulting in considerable deviation from the crystal structures on minimization. This and the distortion introduced on mimization of complexes with significant steric overlap may account for the larger error of prediction seen with the thermolysin test series. As was discussed earlier, an attempt to incorporate explicit hydrogen bond information did not improve predictability. This is not surprising when one considers the very limited correlation of

the electrostatic interaction energy. Atom-centered point charges and hydrogen-bonding information simply may be too crude a representation to give a reliable estimate for this field. In an attempt to test this possibility, we are investigating the use of *ab initio*-derived multipoles on both the receptors and ligands.

Steric fit is another field which did not make a significant contribution to the current PLS model. This field has had an odd history. In the initial stages of development, the steric fit parameter maintained a significant contribution. As the size and sophistication of the training set grew, however, the contribution of this field fluctuated up and down significantly. Although the removal of this field is under consideration, we would like to see the contribution consistently remain low as the size of the training set increases before doing so. It is possible that the information contained in this field is redundant since we already compute steric interaction energy. Currently, there are 51 complexes in the training set. When the complexes in the first test set are added, this number will increase to 65. Our intent is to continue to add complexes of known structure and affinity until the coefficients and contributions of the various parameters converge.

The relatively good predictions of the three DNA-binding molecules in the test set were unexpected considering the apparently small contribution of electrostatics to the model and the fact that these are highly charged systems and no nucleic acid complexes were represented in the training set. No significant conclusions can be drawn from only three observations, and it will be necessary to test several more DNA-binding molecules to verify these results. The excellent performance on the HIV test set was not surprising when one considers that 15 of the complexes in the training set are HIV complexes.

The problems with the five poorly predicted thermolysin inhibitors are likely 2-fold. First, the charge representation issue becomes magnified due to the presence of the bound zinc in the active site of the receptor and the phosphorus-based groups in several of the ligands. The second, and most likely the largest, factor is the alignment and resultant enormous starting energies for the minimization. Significant steric overlaps existed in most of the complexes produced from this alignment which was based on field fit. The starting energies did raise a red flag; however, we felt it would be a good test of the overall process to determine how sensitive the method was to the starting geometry of the complex. To combat this problem, we are currently investigating methods for the flexible docking of ligands to the receptor. The HIV test set illustrates that good predictions can be made for systems in which the crystal structure of the receptor−ligand complex is not known. However, predictions from systems in which the optimal ligand−receptor interaction is not clear are highly suspect.

The results of the new method VALIDATE are encouraging as the optimal set of relevant parameters have yet to be determined. Perhaps more encouraging is the fact that we may be able to significantly improve the model with an improved representation of electrostatics and metal parameterization. Even with current limitations, VALIDATE provides a reasonable method for the estimate of the affinity of ligands with a known receptor. This is true whether the receptor has a contingent of ligands with known binding affinities or not. Reliable predictions of affinity are crucial to the success of rational drug design, protein engineering, and understanding of protein−protein interaction; we believe that VALIDATE is a positive step in that direction.

JA9539002