# Three-Dimensional Quantitative Structure–Activity Relationship of Human Immunodeficiency Virus (I) Protease Inhibitors. 2. Predictive Power Using Limited Exploration of Alternate Binding Modes

Tudor I. Oprea,[†] Chris L. Waller,[‡] and Garland R. Marshall*

*Center for Molecular Design, Washington University School of Medicine, Lopata Hall, Box 1099, 1 Brookings Drive, St. Louis, Missouri 63130-4866*

*Received January 31, 1994*[⊛]

NewPred, a semiautomated procedure to evaluate alternate binding modes and assist three dimensional quantitative structure–activity relationship (3D-QSAR) studies in predictive power evaluation is exemplified with a series of 30 human immunodeficiency virus 1 protease (HIV PR) inhibitors. Five comparative molecular field analysis (CoMFA) models (Waller, C. L.; *et al. J. Med. Chem.* **1993**, *36*, 4152–4160) based on 59 HIV-PR inhibitors were tested. The test set included 18 compounds (set A) having a different transition state isostere (TSI), hydroxyethylurea (Getman, D. P.; *et al. J. Med. Chem.* **1993**, *36*, 288–291), to investigate the binding mode in P1′ and P2′. Twelve dihyroxyethylenes (set B) (Thaisrivongs, S.; *et al. J. Med. Chem.* **1993**, *36*, 941–952) were used to investigate binding in P2 and P3 as well as in P2′ and P3′. Six other compounds with known or inferred binding structure (set C) were part of the test set, but not investigated with NewPred. Each compound was aligned in accordance to predefined alignment rules for the training set prior to the inclusion in the test set (except for set C). Using NewPred, geometrically different conformers for each compound were generated and individually relaxed in the HIV-PR binding site. Energy comparisons allowed selection of lowest energy structures to be included in the test set. Only *in vacuo* minimized conformers derived from low-energy complexes were used to determine the predictive power of the five models (predictive $r^2$ varied from 0.1 to 0.7 when two chemical and statistical outliers were excluded). Our models correctly predict the poor inhibitor activity of 1(*S*)-amino-2(*R*)-hydroxyindan-containing peptides (set B), which is explained and interpreted from a 3D-QSAR perspective. The use of a new, flexibility-based, semiautomated method to explore alternate binding modes for 3D-QSAR models is demonstrated.

## Introduction

Comparative molecular field analysis[1] (CoMFA) is a three-dimensional quantitative structure–activity relationship (3D-QSAR) approach[2] that computes the steric and electrostatic interactions of a given series of molecules with a regular lattice of probe atoms.[1,3] The quantitative results are tabulated, and appropriate statistic techniques yield an equation (QSAR) outlining the key features of the model that explain variability in the target property based on variation in the molecular fields of the studied compounds. Recommended statistical techniques for CoMFA studies are partial least squares[4] (PLS) and principal component analysis[5] (PCA), with cross-validation to select, among several PLS models, the one with the highest predictive value.[6]

When analyzing ligands to generate a 3D-QSAR model, flexible compounds are by far the most difficult. A choice for the active conformation for each molecule and the corresponding superposition have to be generated, either in accordance with available experimental data or based on hypothetical assumptions. Thus, one of the key steps in 3D-QSAR methodology is selection of the conformation for each ligand in the series, followed by molecular superposition (alignment rules). The underlying success of a 3D-QSAR model is dependent on both decisions.

To define the alignment rules for a flexible training set, one can use a variety of methods. If crystallographic data are available, the field-fit alignment procedure[7] may prove useful (crystals being used as template molecules). The field-fit procedure minimizes the RMS difference between a fixed (steric and electrostatic) template field and the corresponding fields of the molecules being aligned by adjusting atomic coordinates (hence, field values). This procedure has been extensively discussed[8] in conjunction with alignment issues and applied to determine the alignment of 52 human immunodeficiency virus 1 protease (HIV PR) inhibitor peptides[9] based on 7 experimentally determined structures of inhibitor–enzyme complexes.

When no structural data are available, methods that investigate conformational space (e.g., using simulated annealing and cluster analysis[10]) may find the best match between various ligands. During this procedure,[10] low-energy conformers are selected and minimized pairwise, and the best match obtained from all different conformations can be selected. This method is useful when no crystal data are available, for structurally dissimilar ligands.

For unknown receptor sites, the active analog approach may be used in conjunction with (constrained) systematic search[11] (implemented as the RECEPTOR module[12] in Sybyl) to generate a set of sterically allowed conformations and to determine the existence of common 3D orientations of specified functional groups, or active site points, in a series of compounds (i.e., the pharmacophore).

For pharmacophoric pattern identification, an automated procedure, DISCO,[13] can be used to generate several

* To whom correspondence should be addressed.

[†] Permanent address: Department of Physiology, Section Molecular Physiology, University of Medicine and Pharmacy, P-ta E. Murgu nr.2, R-1900 Timisoara, Romania.

[‡] Current address: Pharmacokinetics Branch (MD-74), Environmental Toxicology Division, Health Effects Research Laboratory, United States Environmental Protection Agency, Research Triangle Park, NC 27711.

[⊛] Abstract published in *Advance ACS Abstracts*, June 1, 1994.

0022-2623/94/1837-2206$04.50/0

pharmacophoric maps. Each of these represents in itself a possible alignment rule and can be used to generate a 3D-QSAR (CoMFA) model. DISCO includes a suite of programs that allows input or selection of low energy conformations for the compounds to be compared and superposition processing (with ALADDIN[14]). This procedure is particularly useful when large numbers of structurally different compounds are to be investigated.

The fundamental problem of the above-mentioned (and similar) methods is that the proposed solution is often not unique. For flexible molecules, many conformers can match a particular pharmacophoric pattern, and the rationale for choosing one (the "alignment rule") is usually done on an energetic basis. When the choice of the alignment has no reference to experimentally determined structures, results have to be treated with caution,[8] because other conformations may in fact bind to the receptor, and the correlations obtained from the proposed alignment rule may be spurious (compensating inadequacies in the calculation of entropic and enthalpic effects). A similar problem is encountered in choosing test set conformers.

To assist selection among various 3D-QSAR models, an external set of compounds with known activities not used in model generation (referred to as the external or test set) is usually predicted. When a test set is generated, a different situation occurs: the training set (molecules included in the 3D-QSAR model) has been obtained, and a set of alignment rules exists. The alignment for the test set molecules is implicitly constrained by the existing model. Applying the same conformational choices and superimposition procedures to generate a test set of single conformers is useful only if insignificant changes exist in the structures present in the test set, compared to molecules in the training set.

The appropriate conformation for test set molecules is ambiguous even within the alignment rules, if they have flexible moieties not present in the training set. In this case, a (limited) conformational analysis performed on the test set molecules, using the alignment rules as constraints during conformational search, generates multiple conformers for the same ligand. All conformers are consistent with the initial model yet geometrically different and hence with a range of predicted activities that often spans several log units, instead of a single value. A set of 12 conformers of compound M3 in this study are shown as example: all conformers were obtained by active site minimization and are consistent with the alignment rules, yet geometrically different. Their activity was predicted using alignment IV and varied between 0.08 and 1.68 log units (see Figure 1). All these conformers represent theoretical solutions of the conformation achieved in the binding site. While their activity spans 1.6 log units (2.14 kcal/mol in terms of binding affinity), the total energy of the complex (binding site and ligand) spans 330 kcal/mol. Selection of the active conformation among multiple computed possibilities for the same ligand was rationalized on the basis of the calculated energy of the entire complex.

On the basis of this observation, we propose a semiautomated procedure, compatible with the Sybyl/CoMFA method, NewPred. This procedure allows limited conformational analysis based on the alignment rules of an initial CoMFA model, automatically selects a single conformation for test set compounds, and then predicts activities based on the initial QSAR. All conformers are minimized, either in the average steric and electrostatic field of the CoMFA model (option available in Sybyl) using the field fit procedure or, preferably, in the receptor-binding site (when available) to optimize individual conformer alignment. The lowest energy conformer found is then chosen to be included in the final test set, which is used to evaluate the predictive power of the model.

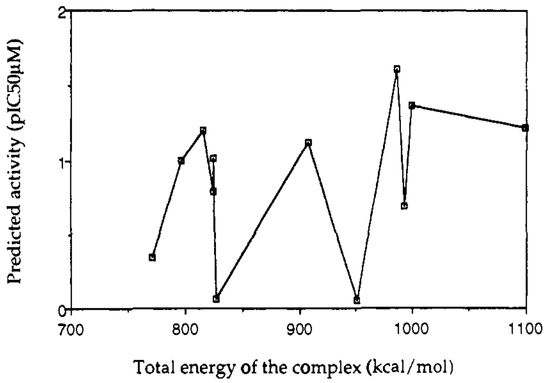


**Figure 1.** Total energy of active site and inhibitor complex vs predicted activity of isolated conformers for compound M3. Twelve local minima, representing geometrically and biologically different conformers, were obtained using NewPred (see Table 2 for the active torsion angles). The activities were predicted using alignment IV.

NewPred was tested on a set of five CoMFA models for human immunodeficiency virus 1 protease (HIV-PR) inhibitors,[9] based on 59 compounds (the training set) representing five different transition-state isosteres (TSI). NewPred was used to select a unique conformation for 30 of the 36 compounds (including three crystal structures and two different TSI classes[15,16]) selected for the test set. Activities were predicted with each of the five models, and the procedure was evaluated in terms of predictive power. On the basis of the proposed conformations, the poor activity of a series of 1(*S*)-amino-2(*R*)-hydroxyindan-containing peptides[17] was explained.

## Methods

**A. Molecular Modeling Methods.** All calculations were performed in Sybyl[18] using the standard Tripos force field.[19] *In vacuo* minimizations were performed with an energy change convergence criterion of 0.001 kcal/mol. In this study, *in vacuo* minimization refers to ligand minimizations performed in the absence of the active site. Minimizations *in situ* (which refer to minimizations performed in the presence of the active site) used an energy change criterion of 10 kcal/mol during the limited conformational analysis and 5 kcal/mol when dihedral angles were modified with a small angle increment. These energy criterions were used to save CPU time because a large number of conformers (up to 500) were minimized in the active site. The active site was defined as a substructure sphere of 12 Å radius extracted from the Roche inhibitor/HIV PR complex, centered on the hydroxyl oxygen (at the hydroxyethylamine TSI) of the Roche ligand, to avoid time-consuming computations. Backbone atoms of the selected active site were kept rigid during minimization. Side-chain atoms and ligand atoms were allowed to relax. Water oxygens were rigid to conserve the internal hydrogen bonding pattern.

For the active site, partial atomic charges[20] were loaded from the Sybyl Biopolymer dictionary (Kollman all-atom method). Partial charges were calculated using Mopac[21] 5.0 with the AM1 Hamiltonian (Mopac keywords: AM1

1SCF MMOK), for all ligands, as well as for all essential water molecules and nonstandard residues (protonated Asp[125]) in the active site sphere. All energy minimizations and semiempirical calculations were performed on IBM 560 and Silicon Graphics Iris 4D/380 workstations.

**B. 3D-QSAR Methodology.** CoMFA calculations used the following characteristics: the grid was regularly spaced (2 Å), with dimensions of 36 × 26 × 22 Å; steric and electrostatic calculations were performed using a carbon $sp^3$ probe atom with a -1 charge, and a distance-dependent dielectric constant and a cutoff of ±30 kcal/mol, with no electrostatic interactions at steric bad contacts. The same CoMFA grid box was used in all five models and predictions. All CoMFA analyses were computed on a Silicon Graphics Iris 4D/380 computer, with the Sybyl 6.0 package.

Regression analyses were done using the Sybyl implementation of the PLS[4] algorithm, initially with cross-validation (the leave-one-out technique), and 10 principal components (PCs). The optimal number of components to be used in conventional analyses was chosen from the analysis with the highest cross-validated $r^2$ value, and for component models with identical $r^2$ values, the model with the smallest standard error of prediction. To improve the signal-to-noise ratio, all leave-one-out calculations were performed with a 2.0 kcal/mol energy column filter (minimum_sigma, or field variance at each grid point).

The predictive $r^2$ was used to evaluate the predictive power of the CoMFA model, and was based only on molecules from the test set. Predictive $r^2$ is calculated using the formula:

$$\text{predictive } r^2 = 1 - (\text{"press"}/\text{SD})$$

where SD is the sum of the squared deviations between the actual activities of the compounds in the test set and the mean activity of the training set compounds and "press" is the sum of the squared deviations between predicted and actual activities for every compound in the test set. Prediction of the mean value of the training set for every member in the test set yields a predictive $r^2 = 0$, while negative values are possible when the predictions are worse than predicting the mean value of the training set. All predicted activities for the test set molecules were obtained using the CoMFA models established for each alignment (I–V) as presented elsewhere[9] and summarized below.

**C. The Five CoMFA Models.** CoMFA was used to examine the correlation between calculated physicochemical (steric and electrostatic) properties and measured *in vitro* inhibitory activities of a series of 59 HIV-PR inhibitors.[9] Five different TSIs were represented: hydroxyethylamine,[22–25] statine,[26] norstatine,[27] ketoamide,[27] and dihydroxyethylene.[28] Seven crystal structures of inhibitor–protease complexes (Roche,[22] JG365,[15] U75875,[29] Ag1001,[30] Ag1002,[30] Ag1004,[30] and L689,502[31]) provided information regarding active conformations and relative positions of different ligands within the binding site.

These experimentally determined alignment rules were used for the rest of the training set. Field-fit minimization of neutral ligands using the corresponding TSI crystal is referred to as alignment I. Charged species of Alignment I geometries (no minimization, except at the local ionized moieties) constituted alignment II, while a reminimized version of alignment II generated alignment III. Another alignment rule was determined by minimizing each field-fitted ligand in the binding site. Active site minimized compounds derived from Alignment I constituted alignment IV, while charged species (from alignment III) were minimized in the active site to generate alignment V. These five alignment rules were discussed previously,[9] and one of them, alignment I, was given preference based on predictive power evaluation against the same test set (termed set A in this paper).

**Table 1.** Summary of CoMFA Results for Each Alignment, including Predictive $r^2$ for the Entire Test Set ($r^2_{\text{pred}-36}$) and after the Exclusion of Two Outliers ($r^2_{\text{pred}-34}$)[a]

| | alignment rule | | | | |
|---|---|---|---|---|---|
| | I | II | III | IV | V |
| $r^2_{\text{cross}}$ | 0.778(6) | 0.653(8) | 0.607(8) | 0.659(7) | 0.642(7) |
| sep | 0.552 | 0.704 | 0.749 | 0.684 | 0.707 |
| $r^2$ | 0.984(6) | 0.990(8) | 0.991(8) | 0.988(7) | 0.983(7) |
| $s$ | 0.146 | 0.122 | 0.112 | 0.129 | 0.156 |
| $F$-test | 549.838 | 597.130 | 703.933 | 592.234 | 413.512 |
| $r^2_{\text{pred}-36}$ | 0.490 | 0.402 | 0.466 | 0.108 | 0.447 |
| $r^2_{\text{pred}-34}$ | 0.679 | - | - | 0.701 | 0.563 |

[a] The test sets are entirely compatible and corresponding to alignments IV and V only. For comparative purposes, the neutral test set (IV) was predicted with alignment I, while the ionized test set (V) was used for alignments II and III.

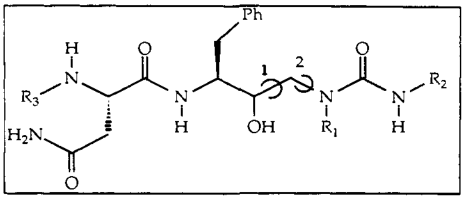
The test set (A) consisted of 18 molecules containing a different TSI, hydroxyethylurea,[16] which were aligned to the Roche crystal, and also in a "flipped" conformation, based on structural details concerning an unexpected binding mode for the Monsanto crystal[16] (flipped in P1′ and P2′, when compared to the Roche crystal). Seven out of 18 compounds were predicted (based on minimization in the binding site) to interact with the enzyme in the Roche-like mode, while the other 11 compounds were found to bind in the flipped Monsanto-like mode.

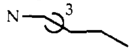
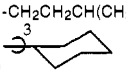
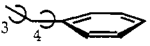
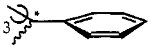
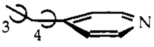
The statistical results (including the predictive $r^2$ values reported in this study) have been duplicated[9] in Table 1 for the convenience of the readers.

**D. The Test Set.** All molecular structures from the test set were built with Sybyl[18] using similar crystal structures as template molecules. Three distinct categories of inhibitors were present in this test set: hydroxyethylureas (Monsanto compounds, referred to as set A), dihydroxyethylenes (Upjohn compounds, referred to as set B), and six other structures (set C). The 18 hydroxyethylureas[16] (set A) were aligned to the Roche crystal, and the binding mode in P1′ and P2′ was investigated using NewPred because of the above-mentioned flipped binding mode.[16] The 12 dihydroxyethylenes (set B) were aligned to the U75875[28] crystal, as they were part of the same TSI class. These structures include very flexible substituents in the P2 and P3 regions, as well as different rigid substituents at the P2′ and P3′ regions. All these regions were explored to determine their probable binding mode. The dihedral angles that were investigated are marked for each compound in sets A (Table 2) and B (Table 3).

The six other structures (set C) used were pepstatin A, acetylpepstatin[32] (containing the statine TSI), MVT101[33] (containing the reduced amide TSI), U85548e[34] (containing the hydroxyethylene TSI), and KNI-93 and KNI-122, containing the norstatine[35] TSI. Pepstatin A was obtained by modifying acetylpepstatin and using information[26] concerning its binding mode, while acetylpepstatin, MVT101, and U85548e were crystal structures. Two compounds containing only moieties present in different training set compounds, KNI-93 and KNI-122, were assembled from those other structures, and their 3D structures were consistent with the training set. These

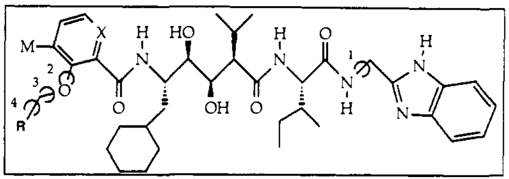
**Table 2.** Structural Formulas, Activities, and Investigated Torsions for Compounds included in Set A (Monsanto Compounds)[a]



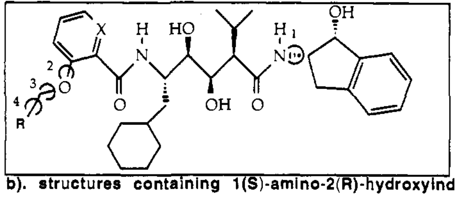
| id | $R_1$ | $R_2$ | $R_3$ | $pIC_{50}(\mu M)$ | Con. |
|---|---|---|---|---|---|
| **M3** | $-CH_2CH(CH_3)_2$ | $-CH_3$ | Cbz | -0.176 | 12 |
| **M4a** | $-CH_2CH(CH_3)_2$ |  | Cbz | 0.026 | 48 |
| **M4b** | $-CH_2CH(CH_3)_2$ | same as **M4a** | Qua | 0.899 | 48 |
| **M5** | $-CH_2CH(CH_3)_2$ | $-(CH_2)_2CH_3$ | Cbz | 0.285 | 12 |
| **M6** | $-CH_2CH(CH_3)_2$ | $-CH_2CH_3$ | Cbz | 0.481 | 12 |
| **M7** | $-CH_2CH(CH_3)_2$ | $-CH(CH_3)_2$ | Cbz | 0.585 | 12 |
| **M8a** | $-CH_2CH(CH_3)_2$ | $-C(CH_3)_3$ | Cbz | 1.456 | 12 |
| **M8b** | $-CH_2CH(CH_3)_2$ | $-C(CH_3)_3$ | Qua | 2.221 | 12 |
| **M9a** | $-CH_2CH_2CH(CH_3)_2$ | $-C(CH_3)_3$ | Cbz | 1.886 | 12 |
| **M9b** | $-CH_2CH_2CH(CH_3)_2$ | $-C(CH_3)_3$ | Qua | 2.523 | 12 |
| **M10a** |  | $-C(CH_3)_3$ | Cbz | 1.537 | 24 |
| **M10b** | same as **M10a** | $-C(CH_3)_3$ | Qua | 2.301 | 24 |
| **M11a** |  | $-C(CH_3)_3$ | Cbz | 1.721 | 48 |
| **M11b** | same as **M11a** | $C(CH_3)_3$ | Qua | 2.523 | 48 |
| **M12** |  | $-C(CH_3)_3$ | Cbz | -0.813 | 36 |
| **M13** | same as **M12** | $-C(CH_3)_3$ | Cbz | -0.707 | 36 |
| **M14a** |  | $-C(CH_3)_3$ | Cbz | 0.978 | 48 |
| **M14b** | same as **M14a** | $-C(CH_3)_3$ | Qua | 1.721 | 48 |

[a] Torsions investigated with NewPred and the number of conformers generated by the initial step are given for each compound. Each torsion angle was investigated with a 60° increment. The rigid moieties were aligned to the Roche crystal.[19] Cbz, carbobenzyloxy; Qua, quinoline-2-carboxamide. Carbon marked (*) in R1 is *R* in M12 and *S* in M13; Con, number of conformers investigated with NewPred; $pIC_{50}$ ($\mu$M) calculation is explained in Appendix.

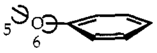
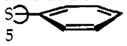
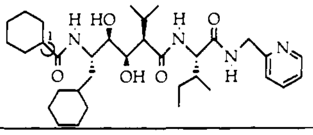
**Table 3.** Structural Formulas, Activities, and Investigated Torsions for Compounds included in Set B (Upjohn Compounds)[a]



a). structures containing 2-(aminomethyl)benzimidazole



b). structures containing 1(S)-amino-2(R)-hydroxyindan

| id | type | R | X | $pIC_{50}(\mu M)$ | Con. |
|---|---|---|---|---|---|
| **U1** | a | --- | CH | 1.092 | 24 |
| **U2** | b | --- | CH | 1.444 | 16 |
| **U8** | a |  | CH | 1.745 | 512 |
| **U9** | a | same as **U8** | N | 0.268 | 64 |
| **U10** | a | same as **U8** | CH | 0.967 | 64 |
| **U16** | b | same as **U8** | CH | 0.268 | 432 |
| **U17** | b | same as **U8** | N | 0.268 | 64 |
| **U18** | a |  | CH | 1.950 | 432 |
| **U19** | a | same as **U18** | N | 0.502 | 64 |
| **U20** | b | same as **U18** | CH | 0.347 | 64 |
| **U21** | b | same as **U18** | N | -0.380 | 64 |
| U-B | |  | | 1.305 | 4 |

[a] The rigid moieties were aligned to the U75875 crystal.[26] Type denotes the nature of the substituent, see drawings a and b. Structures U1 and U2 have no substituent at the phenyl (torsions 2–4 are invalid); the orientation of the phenyl ring was examined instead. M in drawing a denotes a methyl for U10, and a hydrogen for all other structures. Structure U-B is presented entirely, being similar to the U75875 crystal,[26] except for the terminal cyclohexyl (investigated with NewPred). See Table 2 for details concerning torsional investigations and semnification of headings.

six compounds were predicted without conformational analysis. The structures of set C compounds are shown in Figure 2. Data transformation methods for activities of sets A, B, and C are reported in the Appendix. The numbering of compounds included in sets A and B is the same as in the original papers.[16,17]

Alignments IV and V were obtained from active site minimizations and were suitable for NewPred investigation, which generated test set molecules compatible and internally consistent with these two models. Test set IV was used for predictions with neutral alignments (I and IV), while test set V was used for ionized alignments (II, III and V). Predictions for alignments I–III are given for *comparative* purposes only, whereas predictions with models IV and V represent rigurous evaluations.

**E. NewPred.** The semiautomated procedure for limited conformational analysis, NewPred, was written in Sybyl programming language[18] (SPL) and is designed to be used interactively. A flow chart of the program is presented in Scheme 1. The active dihedral angles are user-defined for each compound, and several computational parameters (minimization criteria, dihedral angle range and increment, etc.) can be modified. The program then automatically minimizes each conformer in the binding site, using the Sybyl software. The starting conformation for NewPred has to be compatible with the training set alignment rules (i.e., peptide backbone and sidechains that are encountered in the training set have to be aligned in a consistent manner). The investigated moieties should be allowed to relax to a local minima using a constrained minimization procedure.

Minimizations can be performed either in the average steric and electrostatic field of the training set (these fields can be retrieved from Sybyl/CoMFA and the field-fit procedure is then invoked) or in the ligand-binding site structure (when available). In this study, a 12-Å-radius sphere centered on the binding site was used for minimization purposes, as mentioned before. Results were stored in table format (one table per compound) from which data can be later retrieved. NewPred then selects the five conformers with the lowest energies of the inhibitor-binding site complex.

These five conformers are submitted to another step in conformational analysis when the chosen dihedral angles are rotated up to ±15° with a 3° increment. This allows for fine tuning of the minimization in the active site, since the first minimization step does not explicitly allow flexibility of the ligand. The resulting structures are sorted on a total energy basis, and the conformer that generates the lowest energy enzyme–inhibitor complex is then predicted as such and after *in vacuo* minimization. The *in vacuo* minimization step is required because the binding site has a rigid backbone, which sometimes forces the ligand to distorted geometries (e.g., puckered aromatic rings). In

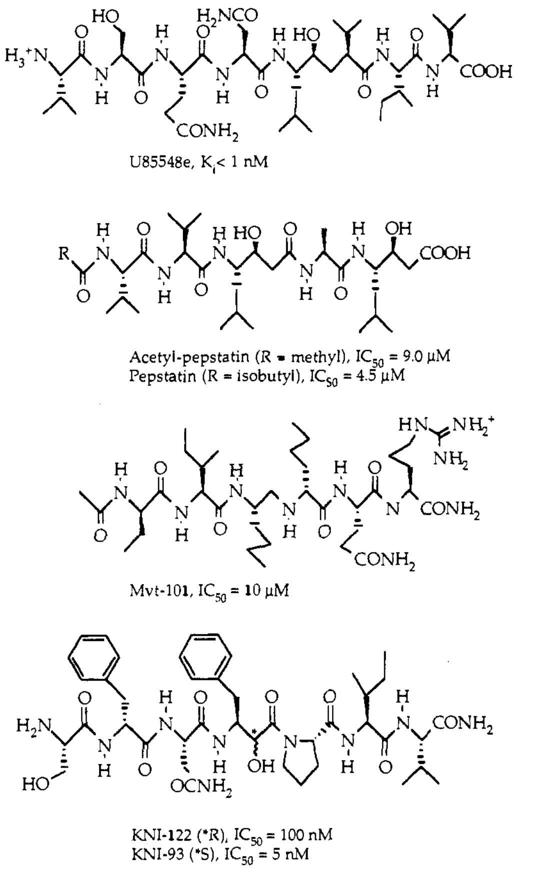

U85548e, $K_i$< 1 nM
Acetyl-pepstatin (R = methyl), $IC_{50}$ = 9.0 μM
Pepstatin (R = isobutyl), $IC_{50}$ = 4.5 μM
Mvt-101, $IC_{50}$ = 10 μM
KNI-122 (*R), $IC_{50}$ = 100 nM
KNI-93 (*S), $IC_{50}$ = 5 nM

**Figure 2.** Structural representation and biological activities for compounds included in set C.

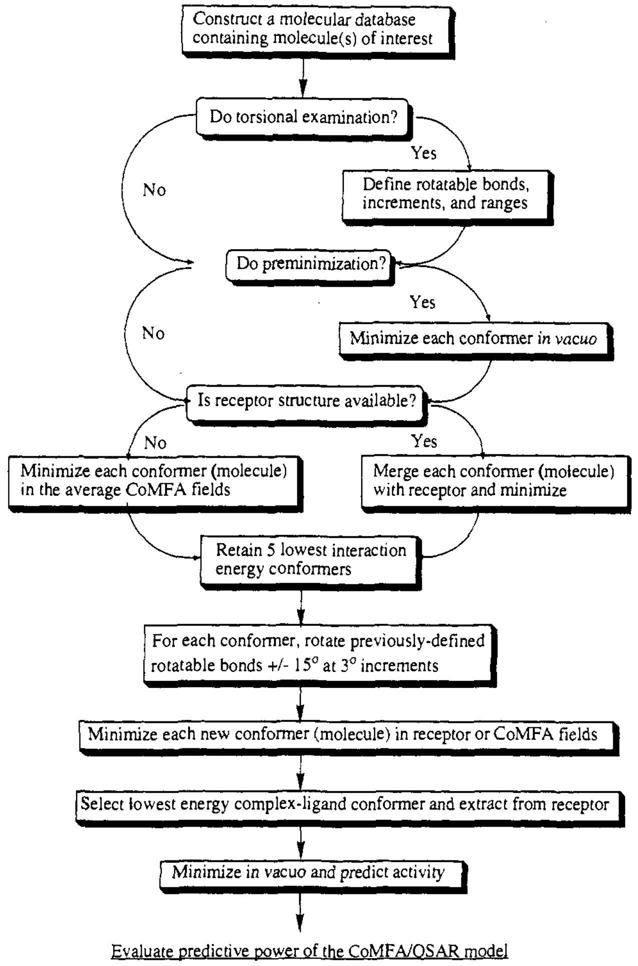
nature, the backbone has a (limited) degree of flexibility to accommodate the ligand in a particular conformation that is not likely to be constrained and possibly similar to the *in vacuo* geometry of the chosen lowest energy conformer. The predicted activity for the *in vacuo* minimized conformer was used for evaluation purposes, and does not differ significantly from the *in situ* minimized structure prediction—up to 0.3 log units (see Appendix for details).

In summary, NewPred consists of four steps: three minimizations, followed by prediction. The first step consists of active site minimization of user-defined conformers (based on choices of active dihedral angles and selected increment steps). In the second step, the five lowest energy complexes are selected, the ligands are extracted and then used to automatically generate new conformers based on the active torsions (±15°, at 3° increment), and each conformer is minimized in the active site. In the third step, the lowest energy conformer is extracted from the binding site and relaxed *in vacuo* to its nearest local minimum. In the fourth step, the conformer's activity is predicted using the available CoMFA model.

## Results and Discussion

**A. Comparison of Predictive Power Results between Various Models.** The results of predictions are presented in Table 4. Initially, they were made for 36 compounds, but two compounds, U85548e and MVT101, were excluded as chemical and statistical outliers (see prediction of set C for details). Therefore, 34 compounds were used to analyze alignments I, IV, and V. Predictive power improved after the exclusion of outliers for all models. Predictive power results were summarized in Table 1.

**Scheme 1.** Flow Chart for the NewPred SPL Procedure As Implemented for the Sybyl/CoMFA Procedure (See Text for Details)


Construct a molecular database containing molecule(s) of interest
Do torsional examination?
Yes
No
Define rotatable bonds, increments, and ranges
Do preminimization?
Yes
No
Minimize each conformer in vacuo
Is receptor structure available?
No
Yes
Minimize each conformer (molecule) in the average CoMFA fields
Merge each conformer (molecule) with receptor and minimize
Retain 5 lowest interaction energy conformers
For each conformer, rotate previously-defined rotatable bonds +/- 15° at 3° increments
Minimize each new conformer (molecule) in receptor or CoMFA fields
Select lowest energy complex-ligand conformer and extract from receptor
Minimize in vacuo and predict activity
Evaluate predictive power of the CoMFA/QSAR model

Alignment I expresses good predictive power for the test set, improved after the exclusion of the outliers. This model has similar predictive power for the Monsanto set ($r^2_{pred}$ = 0.662) and for the final 34 compounds ($r^2_{pred}$ = 0.679), a trend toward underprediction (23 compounds out of 36 were underpredicted) and the smallest average absolute error (0.46 log units or 0.57 kcal/mol).

Alignments II and III expressed poor predictive ability and have a similar tendency to underpredict (21 and 24 underpredicted compounds, respectively). Alignment II has the highest average absolute error, 0.702 log units (0.96 kcal). The largest errors in prediction (in brackets for alignment III) are observed with pepstatin A, 1.65 (1.89) log units, or 2.2 (2.56) kcal/mol, and acetylpepstatin, 1.52 (1.81) log units, or 2.03 (2.42) kcal/mol, respectively.

Alignment IV shows considerable improvement in predictions: from 0.108 on 36 compounds to 0.701 on 34 compounds. This model has the highest errors in prediction for compounds U895548e (4.7 log units, or 6.3 kcal/mol in binding affinity) and MVT101 (2.4 log units, or 3.17 kcal/mol), an average absolute error of prediction of 0.49 log units (0.61 kcal/mol) for the remaining 34 compounds, and a trend toward underprediction (24 compounds out of 36 were underpredicted).

Alignment V shows similar trends with alignments II and III, although the test set was aligned in a consistent

**Table 4.** Differences between Predicted and Actual Activities ($pIC_{50}$, $\mu M$) for the Test Set Molecules

| compound | actual | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| M3 | -0.176 | 0.314 | 0.551 | 0.521 | 0.299 | 0.482 |
| M4a | 0.026 | 0.027 | 0.671 | 0.612 | 0.069 | 0.137 |
| M4b | 0.899 | 0.826 | 0.547 | -0.112 | 0.849 | 0.096 |
| M5 | 0.285 | -0.324 | -0.088 | -0.082 | -0.336 | -0.408 |
| M6 | 0.481 | -0.257 | -0.288 | -0.273 | -0.262 | -0.057 |
| M7 | 0.585 | -0.142 | 0.192 | 0.085 | -0.149 | -0.206 |
| M8a | 1.456 | 1.243 | 0.881 | 0.802 | 1.224 | 0.367 |
| M8b | 2.221 | 0.235 | 0.072 | -0.109 | -0.271 | -0.358 |
| M9a | 1.886 | -0.860 | -0.668 | -0.443 | -0.878 | -0.771 |
| M9b | 2.523 | 0.013 | -0.002 | -0.307 | -0.602 | -0.512 |
| M10a | 1.537 | -1.271 | -1.253 | -0.829 | -1.098 | -1.462 |
| M10b | 2.301 | -1.007 | -0.874 | -0.902 | -0.998 | -0.547 |
| M11a | 1.721 | -0.582 | -0.659 | -0.787 | 0.631 | -0.712 |
| M11b | 2.523 | -0.408 | -1.543 | -0.800 | -0.12 | -0.408 |
| M12 | -0.813 | 0.257 | 0.875 | 0.929 | 0.262 | 0.257 |
| M13 | -0.707 | 0.510 | 1.286 | 1.105 | 0.391 | 0.795 |
| M14a | 0.978 | -0.339 | -0.403 | -0.571 | -0.327 | -0.978 |
| M14b | 1.721 | -0.032 | -0.168 | -0.003 | -0.032 | -0.031 |
| U-B | 1.305 | -0.333 | -0.094 | -0.140 | -0.374 | -0.541 |
| U1 | 1.092 | -0.803 | -0.988 | -0.991 | -0.516 | -1.412 |
| U2 | 1.444 | -1.191 | -1.461 | -1.196 | -0.824 | -1.052 |
| U8 | 1.745 | -1.171 | -1.447 | -1.462 | -0.567 | -1.235 |
| U9 | 0.268 | -0.283 | -0.268 | -0.284 | -0.351 | -0.627 |
| U10 | 0.967 | -0.571 | -0.730 | -0.782 | -0.310 | -1.339 |
| U16 | 0.268 | -0.021 | -0.162 | -0.150 | 0.638 | 0.347 |
| U17 | 0.268 | -0.194 | 0.080 | 0.069 | -0.077 | 0.018 |
| U18 | 1.950 | -0.917 | -1.256 | -1.419 | -0.659 | -1.964 |
| U19 | 0.502 | -0.458 | 0.134 | -0.013 | -0.049 | -0.351 |
| U20 | 0.347 | -0.407 | -0.239 | -0.258 | -0.096 | 0.061 |
| U21 | -0.380 | 0.342 | 0.743 | 0.529 | -0.219 | 0.806 |
| acetylpepstatin | -0.954 | 0.346 | 1.652 | 1.889 | 0.963 | 1.585 |
| pepstatin A | -0.672 | 0.085 | 1.516 | 1.807 | 0.762 | 1.567 |
| MVT101 | -0.230 | 0.959 | 1.094 | 1.192 | 2.372 | 0.552 |
| U85548e | 2.745 | -2.980 | -1.086 | -0.478 | -4.707 | -0.987 |
| KNI-122 | 1.000 | 0.395 | 0.633 | 0.683 | -0.252 | 0.167 |
| KNI93 | 2.301 | -0.556 | -0.670 | -0.455 | -0.903 | -0.068 |
| Average Absolute Errors of Prediction | | | | | | |
| for 36 compds | | 0.570 | 0.702 | 0.641 | 0.659 | 0.646 |
| for 34 compds | | 0.460 | - | - | 0.489 | 0.638 |

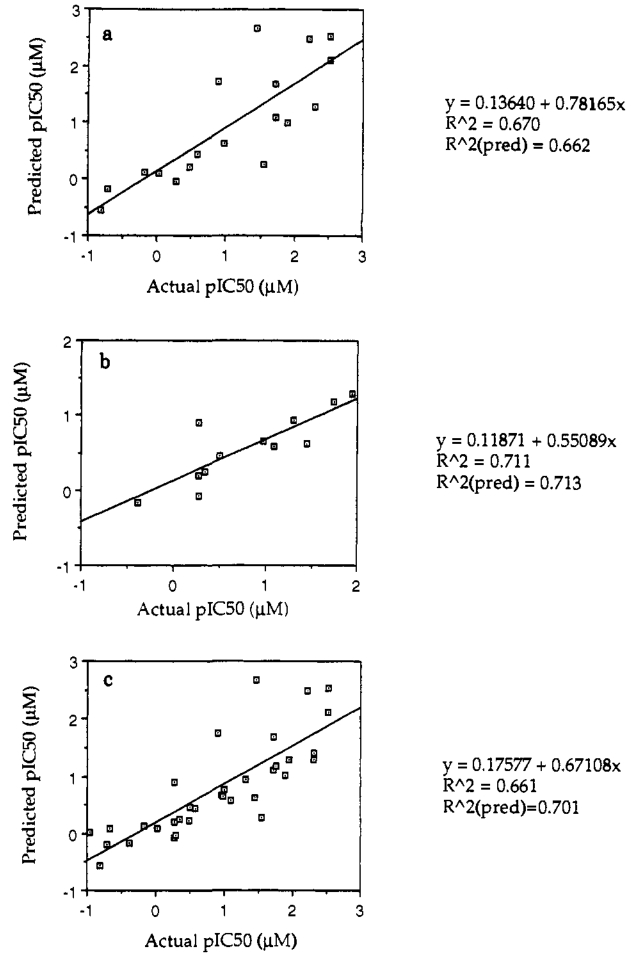


**Figure 3.** Actual vs predicted activities for set A (top), set B (middle), and the entire test set (bottom) using the alignment IV model to generate predictions; the corresponding curve fit equations and correlation coefficients, as well as the predictive $r^2$ are given for each plot.

manner. The results for 36 compounds show a model with predictive ability comparable to alignments I–III and with the same trend toward underprediction (22 underpredicted compounds). Acetylpepstatin and pepstatin A were the highest mispredicted compounds: 1.58 log units (2.12 kcal/mol) and 1.57 log units (2.1 kcal/mol), respectively. The average absolute error was 0.638 log units (0.85 kcal/mol), while the $r^2_{pred}$ improved from 0.447 to 0.563 when the two outliers were excluded.

Assumptions about the ionization state may have been incorrect because various pH assay conditions were used for different classes of compounds; therefore, modeling of charged compounds may have not been accurate (ionization states were not identical for all ligands). Dielectric and, therefore, $pK_a$'s of ionized groups may also change upon binding. However, alignments II, III, and V yielded good regression results; therefore, the lack of predictive power may be due to the composition of the test set. For example, the two highly overpredicted compounds, acetylpepstatin and pepstatin A, were both negatively charged, but no anions were present in the training set—which indicates limited predictive power for models II, III, and V. It has to be noted that none of the Monsanto compounds was ionized, but this did not affect prediction of set A compounds.[9]

The possibility that the good prediction power of neutral models is due to internal consistency with the training sets for alignments I and IV can be ruled out because the test set was consistent with alignment IV only. All further discussions concerning conformer predictions are based on alignment IV, which has the highest predictive $r^2$ when the two outliers have been excluded.

**B. Conformational Analysis and Predictions for Set A.** Although active conformations for each of these 18 molecules were investigated and reported previously[9] in order to establish the binding mode in the P1′ and P2′ sites, a more systematic analysis was performed for the same purpose using the NewPred procedure. The active torsions are shown in Table 2, and the steps performed on each conformer are those from Scheme 1. A plot of actual vs predicted activities is shown in Figure 3a. Lower energy complexes were found than those previously employed in the test set; however, no significant changes in the previously proposed[9] binding modes were observed (data not shown). It is interesting to note that predictions for alignment IV improved significantly for set A, compared to our previous results:[9] $r^2_{pred}$ increased from 0.327 to 0.662—most likely due to the second step in the NewPred analysis, which gives the conformational search a more systematic character. Predictive $r^2$ for set A using alignment I did not change significantly (0.667).

The results of the NewPred procedure have been compared with the experiment: the proposed conformer for compound M4a has a predicted activity of 0.018 $\mu M$ (0.126 $\mu M$ actual $IC_{50}$), and the RMS deviation between the active site minimized structure and the crystallized inhibitor is 0.740 Å. The prediction of the crystal conformation was better, 0.113 $\mu$M, which suggests that the NewPred procedure cannot (and should not) substitute for experimental data.

**C. Conformational Analysis and Predictions for Set B.** The 12 compounds included in set B were selected from a batch of 22 inhibitors recently reported.[17] Selection criteria were (1) flexibility of the compounds (those with more than six potentially active dihedrals were not examined due to the combinatorial nature of the problem, e.g., compounds U3–U7 and U11–U15 have polyethoxy groups[17] that are extremely flexible and difficult to investigate) and (2) the presence in the selected compounds of chemical groups not previously included in the training set or in set A. These structures were aligned to the U75875 crystal,[28] as they have structural similarity and the same TSI. Biological data were originally reported[17] as $K_i$ values, but all data were transformed in $IC_{50}$ values (see Table 3) based on the Cheng and Prussof equation[36] (see Appendix). A plot of actual vs predicted activities is shown in Figure 3b.

The analysis of NewPred-selected conformations in the active site allows several observations concerning structure–activity relations. Examining the binding mode in P2′ and P3′, we noticed that all peptides containing the 2-(aminomethyl)benzimidazole group (U1, U8, U9, U10, U18, and U19) bind with this substituent in the P3′ site, with the NH moiety (benzimidazole) always facing the "back" wall (Gly$^{148}$, Ile$^{150}$, Pro$^{81}$) and hydrogen bonding to one of the crystal water molecules. This binding mode is in agreement to the U75875 crystal data, as modeled in a previous study.[17]

In the same study, the authors state:[17] "We are unable to provide an explanation from the molecular modeling study, however, why the 1(*S*)-amino-2(*R*)-hydroxyindan-containing peptides show much poorer binding affinity to the enzyme." On the basis of NewPred results, all peptides containing the 1(*S*)-amino-2(*R*)-hydroxyindan group (U2, U16, U17, U20, and U21) bind with this substituent located in P2′, yet the long axis of this moiety does not match the long axis of P2′, but is perpendicular. Due to steric interactions, this group cannot be accommodated in the long axis without increased energetic costs; hence, all low-energy conformers of these peptides place the 1(*S*)-amino-2(*R*)-hydroxyindan group parallel to P2′, in an intermediate position between P2′ and P3′. Failure to occupy the P2′ site and the uncharacteristic interaction with the receptor walls (not beneficial to the QSAR in our CoMFA models) probably contribute to the poor activity of these peptides.

The binding mode of the flexible groups in P2 and P3 is not uniquely defined for these Upjohn compounds. However, a general trend is that the first ring starting from P1 toward P2, which is either phenyl or pyridyl for most of these compounds, occupies the P2 site, while the terminal ring (phenyl) is accommodated in the P3 site. Due to the flexibility of the linkage between these two rings, the orientation and position of the terminal phenyl in P3 are variable. Compounds U9, U16, U17, and U18 accommodated the terminal phenyl ring in the P3 binding site parallel to the naphthoxyacetyl group in the U75875 crystal, while other compounds (U8, U10, U19, U20, and U21) accommodate this ring perpendicular to the crystal arrangement. However, for most of these peptides, the steric bulk in this region does not overlap with the beneficial steric fields defined by the CoMFA model; hence part of the decreased biological activity for set B compounds. The high degree of flexibility of these compounds may also include a negative entropic contribution toward the binding free energy in the overall economy of the process.

Replacing one CH group with a nitrogen (by transforming a phenyl ring into a pyridyl), which is predominantly an electrostatic alteration, systematically decreases the activity ($IC_{50}$), which drops from 0.179 (U8) to 0.540 $\mu$M (U9), and from 0.112 (U18) to 0.315 $\mu$M (U19). This trend is reproduced by the model although no pyridyl moieties are located at P2 in the training set. Predicted activities are 0.067 (U8), 1.211 (U9), 0.051 (U18), and 0.353 $\mu$M (U19), respectively. Both the average electrostatic field contributions in the CoMFA model and the U75875 crystal electrostatic field show that in the region occupied by the aromatic nitrogen in the pyridyl-containing peptides, positive charges (e.g., U75875, which has an NH (imidazole) group in P2) or neutral moieties (e.g., in the average CoMFA field obtained from active compounds) are required for good activity. Cross-examination of these compounds in the enzymatic active site shows the carboxyl moiety of Asp$^{30}$ in the near vicinity of the above-mentioned P2 region. These results suggest that insertion of negatively charged groups in P2 is detrimental for biological activity.

**D. Predictions for Set C.** Four compounds for which structural information was obtained from crystallographic studies were added to the test set with the aim to test the predictive power of the models with known active conformations (therefore, no NewPred investigation was required) and to include compounds with other TSIs than already present in the model (e.g., MVT101). The arrangement of these crystals has been recently reviewed.[26] Biological data were adapted from published information concerning U85584e,[37] pepstatin A,[38] acetylpepstatin, and MVT101[39] (see Appendix for details). The proposed active conformations for two other compounds, KNI-93 and KNI-122, were constructed based on fragment similarity with other compounds in the training set and were consistent with our models.

When examining prediction results, out of six compounds, four were predicted within less than 1 log unit from the actual value (acetylpepstatin, pepstatin, KNI-122 and KNI-93), while for two compounds the error was higher than 2 log units (4.71 log units for U85584e and 2.37 log units for MVT101).

The overlap with the CoMFA steric and electrostatic fields is only partial as U85548e extends in regions unoccupied by any structure present in the training set, this compound having two residues outside P3 (occupying P4 and P5 regions). Two residues out of eight are unaccounted for in our models, and they extend at the limit of the CoMFA grid box (see Figure 4), and prediction for those residues is entirely based on extrapolation.

The results observed for MVT101 can be explained by the fact that it has a different TSI (a reduced amide, possibly protonated) and a positively charged residue (Arg) in P2′. This feature is not present in any of the training set compounds and also not in the other compounds included in the test set. A PCA study performed for all 95 compounds (alignment IV) using the GOLPE[40] program shows that, in the first four principal component plots, MVT101 and U85548e are isolated points in cluster space.[41]

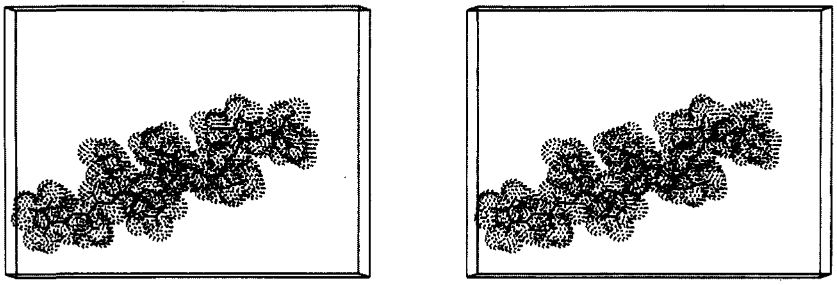


**Figure 4.** The crystal structure of inhibitor U85548e, as aligned in the binding site, is shown in the CoMFA grid box (stereoview, hydrogens omitted for clarity). The P4 and P5 binding sites were not present in any compound from the training set; hence one end of the inhibitor (its van der Waals surface) is at the border of the CoMFA box. This CoMFA box was generated overlapping all training set compounds (alignment I) by at least 4 Å along all axes.

These observations suggest that MVT101 and U85548e are chemical and statistical outliers. Therefore, they were excluded from predictive power evaluations. It is most likely that such outliers exist for each alignment and chemometric tools should be used for this investigation.[41] These results clearly indicate the limitations of the predictive power of these CoMFA models and suggest that careful evaluation of 3D-QSAR models has to be performed to determine predictive power limits.

**E. On the Choice of a Test Set.** The PLS technique raised controversial questions[42] concerning the validation of QSARs, the descriptor-variables pool size, and, implicitly, the trustworthiness of the method. The usefulness of cross-validation, besides the multiple correlation coefficient (conventional $r^2$) and small residual standard deviation, to judge validity of QSARs has been shown.[43] The probability of chance correlation[44] using PLS was recently examined[45] for random data and CoMFA field descriptors, and it was concluded that for data sets with more than 12 compounds, any cross-validated $r^2$ greater than 0.25 from CoMFA is not the result of chance correlation. However, as observed[9] with our five consistent CoMFA models having cross-validated $r^2$ of 0.586 (7) to 0.786 (7), the predictive $r^2$ varied from 0.188 to 0.624 on the same test set, suggesting that beyond the risk of chance correlation, several models should be tested before selecting one QSAR model.

The choice of an external test set for predictive power evaluation of QSARs has to take into account several factors. First, biological assay methods used for the test set should be compatible with those from the training set. Therefore, the selection of compounds to be included in a QSAR model depends on the availability of data from the same laboratory or from laboratories that use compatible (comparable) assay methods. In this study, the selection of the test set was made primarily on the basis of this criterion, and for this purpose the biological data for set B were transformed by using enzyme characteristics reported previously[28] by the same group.

Second, a good test set should span several orders of magnitude in activity, yet not exceeding activity values in the training set by more than 10%, because predictive power should be tested on activities in the range of the training set and activity should not be extrapolated. Extrapolation procedures should be tested separately, since extremely (in)active compounds are likely to be outliers when compared to the training set. Volume (tested compounds have to fit in 3D space defined by the CoMFA model[46]) and alignment compatibility have to be examined, while structural variations in regions where the training set has conserved moieties have to be treated with care.

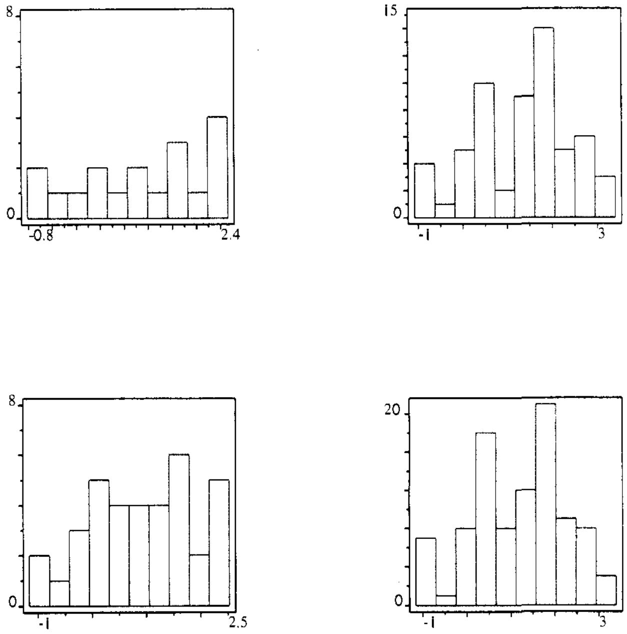


**Figure 5.** Histograms of activity vs number of compounds for set A (top left), training set (top right), the entire test set (bottom left), and the combined training and test set (95 compounds, bottom right).

Third, a balanced test set (in our definition) should also have similar number of (in)active compounds to ensure the uniform sampling of biological activity. In terms of activity range, the training set has a mean activity ($A$) of 0.997 with a standard deviation (stdev) of 1.229, a maximum activity ($A_{max}$) of 3.398, and a minimum activity ($A_{min}$) of −1.301. Among the predicted sets, set A has the following: $A = 1.080$, stdev = 1.072, $A_{max} = 2.523$ and $A_{min} = -0.813$, while the entire test set has $A = 0.472$, stdev = 1.167, $A_{max} = 2.523$ and $A_{min} = -1.381$. For this reason, set A is better as a test set than the union of sets A, B, and C, and its activity values are more evenly distributed (the set is more balanced). The main reason is that sets B and C contain mostly inactive compounds, creating a certain trend in the test set (see histograms in Figure 5).

The bias created by this situation may not be apparent, but the following hypothetical example is illustrative: a QSAR with significant statistical correlation (validated by high $r^2$ values, both cross-validated and conventional) predicts correctly inactive compounds, yet fails to predict active compounds. For a test set containing mostly inactive compounds, the predictive $r^2$ of this model will be close to 1, and its predictive power will be appreciated as good. If active compounds dominate a test set, the predictive $r^2$ will be close to 0, and the model will be rejected. A balanced test set is, therefore, needed to evaluate its predictive power correctly.

The initial test set (now, set A) was reasonably balanced, but predictability concerning other binding pockets except P1′ and P2′ was not tested; hence, we introduced compounds from set B. The ability of these models to predict fixed conformations (crystal structures of different structural classes of inhibitors) was tested with set C. The inclusion of more active and moderately active compounds is clearly required to obtain a balanced test set. The validation of a QSAR should be made on the basis of its predictive power when faced with different compounds (not just with analogs of training set members), first with an internal set (e.g., the leave-five-out technique[40]) to optimize the regression model and later with an external and balanced test set.

## Conclusions

This study addressed a fundamental issue in 3D-QSAR studies: that predictions are evaluated based on the choice of a single (arbitrary) conformation for test compounds, among multiple possibilities. Our proposed method, NewPred, allows the computational exploration of alternate binding modes and the proposed single conformer is the result of a systematic search, within the limits of the initial alignment rules. In this paper, NewPred was tested for a series of 30 flexible HIV-PR inhibitors, and for each of them a single conformation was used for prediction. The possibility to use multiple conformers for the same compound instead of a single conformer is currently under examination. In this case, the predicted activity would become a *range* instead of a single value, perhaps increasing the probability of correctitude.

From the resultant predictive $r^2$ (Table 1), the model of choice is neutral (alignments I and IV), both models correlating better molecular field variance with the biological activity. Because no significant difference in terms of predictive power exists between alignments I and IV, none of these models can be preferred. However, the ionized models (II, III, and V) were less performant. The use of alignment V to evaluate ionized compounds cannot be discarded, as it proved to be marginally predictive. The agreement between the experimentally observed and calculated conformation of compound M4a favor the use of NewPred in the absence of structural data.

When selecting candidates for the test set, care should be exercised to build a balanced test set (containing similar numbers of active and inactive compounds, with compatible biological activities) before evaluating the predictive power of the model.

When multiple conformers of test set compounds are possible and are consistent with the alignment rules, the NewPred procedure may prove to be a useful tool for test set composition and 3D-QSAR model selection. After selecting between several models, an emergent (trustworthy) 3D-QSAR model should be verified experimentally against compounds designed and predicted based on that model. Limitations of predictive power (e.g., extrapolation, unaccounted molecular size, and/or chemical features) due to the inherent bias of the training set should be expected.

**Acknowledgment.** The authors wish to thank Alexander Wlodawer (National Cancer Institute) and William Stallings (Monsanto) for supplying the crystal coordinates of inhibitor/protease complexes. The authors gratefully acknowledge support for this research from National Institutes of Health Research Cooperative Agreement Grant No. 5 UO1 AI27302 and Grant GM 24483. One author (C.L.W.) wishes to acknowledge support from NIH Cardiovascular Training Grant No. T32HLO7275.

**Supplementary Material Available.** Cartesian coordinates for NewPred selected conformers of test set compounds and the NewPred SPL script are available from the authors upon request (email: garland@wucmd.wustl.edu).

## Appendix

Biological activities used in the CoMFA model and predictive power analysis are expressed as

$$\mathrm{pIC}_{50} = -\log_{10} \mathrm{IC}_{50} \qquad (1)$$

where $\mathrm{pIC}_{50}$ is the transformed activity, and $\mathrm{IC}_{50}$ is the micromolar concentration of the inhibitor producing 50% inhibition of the HIV-1 protease substrate cleavage activity.

Biological activities for the peptide series shown in Table 3 were published as $K_i$ values.[17] In order to obtain $\mathrm{pIC}_{50}$ values, the Cheng and Prusoff equation[36] determined for the case involving one substrate and one competitive inhibitor present was used

$$\mathrm{IC}_{50} = K_i\,(1 + S/K_m) \qquad (2)$$

where $K_i$ is the dissociation constant of the enzyme–inhibitor complex, $S$ is the substrate concentration, and $K_m$ is the Michaelis constant of the substrate. This equation is valid when the velocity in the presence of the inhibitor is half the velocity in the absence of the inhibitor.

Based on eq 2, $\mathrm{IC}_{50}$ values for all compounds in the set B were determined using the $K_m$ value of 2.0 mM (Tomasselli, A. G., personal communication), $S = 2.5$ mM, and the corresponding $K_i$ values.[17]

For compound U85584e, data were extrapolated from a plot where the maximum velocity of HIV-1 protease was evaluated with U85548e at three different substrate concentrations[37] and an $\mathrm{IC}_{50}$ of 39 ± 2 nM was obtained. For acetylpepstatin and pepstatin, $\mathrm{IC}_{50}$ values were obtained from D. P. Getman (Monsanto). For MVT101, the $\mathrm{IC}_{50}$ value is available from the literature.[15]

A simple statistical analysis was undertaken to compare differences in predicted activity between *in situ* minimized conformers and their *in vacuo* minimized correspondents. The difference was less than 0.3 log units when comparing large numbers of conformers. For compound U8, the difference in predicted activity ($\Delta p$) was 0.279, with a standard deviation (sdev) of 0.157, a mean absolute deviation (adev) of 0.129 for $n = 512$ conformers; for compound U16, $\Delta p = 0.150$, sdev = 0.138, adev = 0.097, $n = 0.432$; for compound U18, $\Delta p = 0.265$, sdev = 0.202, adev = 0.164, $n = 432$. For the NewPred selected conformers of the combined sets A and B, $\Delta p = 0.117$, sdev = 0.092, adev = 0.067, $n = 30$.

## References

(1) Cramer, R., III; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(2) Marshall, G.; Cramer, R., III Three Dimensional Structure-Activity Relationships. *Trends Pharmacol. Sci.* **1988**, *9*, 285–289.

(3) Cramer, R., III; Bunce, J. The DYLOMMS method: Initial results from a comparative study of approaches to 3D QSAR. In *QSAR in Drug Design and Toxicology*; Hadzi, D., Jerman-Blazic, B., Ed.; Elsevier: Amsterdam, 1987; pp 3–12.

(4) Stahle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196.

(5) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

(6) Cramer, R.; Bunce, J.; Patterson, D.; Frank, I. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.

(7) Clark, M.; Cramer, R. I.; Jones, D.; Patterson, D.; Simeroth, P. Comparative molecular field analysis (CoMFA). 2. Toward its use with 3D-structural databases. *Tetrahedron Comput. Method.* **1990**, *3*, 47–59.

(8) Klebe, G.; Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70–80.

(9) Waller, C.; Oprea, T.; Giolitti, A.; Marshall, G. 3-D QSAR of human immunodeficiency virus (I) protease inhibitors. I. A CoMFA study employing experimentally-determined alignment rules. *J. Med. Chem.* **1993**, *36*, 4152–4160.

(10) Perkins, T.; Dean, P. An exploration of a novel strategy for superposing several flexible molecules. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 155–172.

(11) Dammkoehler, R.; Karasek, S.; Shands, E.; Marshall, G. Constrained Search of Conformational Hyperspace. *J. Comput. Aided Mol. Des.* **1989**, *3*, 3–21.

(12) RECEPTOR, 2.4, available from Tripos Associates, 1699 S. Hanley Rd., St. Louis, MO 63144.

(13) Martin, Y.; Bures, M.; Danaher, E.; DeLazzer, J. New strategies to improve the efficiency of the 3D design of bioactive molecules. In *Trends in QSAR and Molecular Modelling 92*; Wermuth, C., Ed.; ESCOM: Leiden, 1993; pp 20–27.

(14) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: An integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric aand substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 225–251.

(15) Swain, A.; Miller, M.; Green, J.; Rich, D.; Schneider, J.; Kent, S.; Wlodawer, A. X-ray crystallographic structure of a complex between a synthetic protease of Human Immunodeficiency Virus 1 and a substrate-based hydroxyethylamine inhibitor. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 8805–8809.

(16) Getman, D.; DeCrescenzo, G.; Heintz, R.; Reed, K.; Talley, J.; Bryant, M.; Clare, M.; Houseman, K.; Marr, J.; Mueller, R.; Vazquez, M.; Shieh, H.-S.; Stallings, W.; Stegeman, R. Discovery of a Novel Class of Potent HIV-1 Protease Inhibitors Containing the (R)-Hydroxyethylurea Isostere. *J. Med. Chem.* **1993**, *36*, 288–291.

(17) Thaisrivongs, S.; Turner, S.; Strohbach, J.; TenBrink, R.; Tarpley, W.; McQuade, T.; Heinrickson, R.; Tomasselli, A.; Hui, J.; Howe, W. Inhibitors of the protease for the HIV: Synthesis, enzyme inhibition and antiviral activity of a series of compounds containing the dihydroxyethylene transition-state isostere. *J. Med. Chem.* **1993**, *36*, 941–952.

(18) SYBYL molecular modeling system, available from Tripos Associates, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

(19) Clark, M.; Cramer, R., III; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(20) Weiner, S.; Kollman, P.; Nguyen, D.; Case, D. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **1986**, *7*, 230–252.

(21) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.

(22) Krohn, A.; Redshaw, S.; Ritchie, J.; Graves, B.; Hatada, M. Novel Binding Mode of Highly Potent HIV-Proteinase Inhibitors Incorporating the (R)-Hydroxyethylamine Isostere. *J. Med. Chem.* **1991**, *34*, 3340–3342.

(23) Rich, D.; Sun, C.-Q.; Prasad, J.; Pathiasseril, A.; Toth, M.; Marshall, G.; Clare, M.; Mueller, R.; Houseman, K. Effect of hydroxyl group configuration in hydroxyethylamine dipeptide isosteres on HIV protease inhibition. Evidence of multiple binding modes. *J. Med. Chem.* **1991**, *34*, 1222–1225.

(24) Roberts, N.; Martin, J.; Kinchington, D.; Broadhurst, A.; Craig, J.; Duncan, I.; Galpin, S.; Handa, B.; Kay, J.; Krohn, A.; Lambert, R.; Merrett, J.; Mills, J.; Parkes, K.; Redshaw, S.; Ritchie, A.; Taylor, D.; Thomas, G.; Machin, P. Rational design of peptide-based HIV proteinase inhibitors. *Science* **1990**, 258–261.

(25) Tucker, T.; Lumma, W. J.; Payne, L.; Wai, J.; De Solms, S.; Giuliani, E.; Darke, P.; Heimbach, J.; Zugay, J.; Schleif, W.; Quintero, J.; Emini, E.; Huff, J.; Anderson, P. A series of potent HIV-1 protease inhibitors containing a hydroxyethyl secondary amine transition state isostere: synthesis, enzyme inhibition, and antiviral activity. *J. Med. Chem.* **1992**, *35*, 2525–2533.

(26) Appelt, K. Crystal structures of HIV-1 protease-inhibitor complexes. *Perspect. Drug Discov. Des.* **1993**, *1*, 23–48.

(27) Tam, T.; Carriere, J.; MacDonald, I.; Castelhano, A.; Pliura, D.; Dewdney, N.; Thomas, E.; Bach, C.; Barnett, J.; Chan, H.; Krantz, A. Intriguing structure-activity relations underlie the potent inhibition of HIV protease by norstatine-based peptides. *J. Med. Chem.* **1992**, *35*, 1318–1320.

(28) Thaisrivongs, S.; Tomasselli, A.; Moon, J.; Hui, J.; McQuade, T.; Turner, S.; Stronbach, J.; Howe, J.; Tarpley, W.; Heinrikson, R. Inhibitors of the protease from Human Immunodeficiency Virus: Design and modeling of a compound containing a dihydroxethylene isostere insert with a high binding affinity and effective antiviral activity. *J. Med. Chem.* **1991**, *34*, 2344–2356.

(29) Thanki, N.; Rao, J.; Foundling, S.; Howe, W.; Moon, J.; Hui, J.; Tomasselli, A.; Heinrikson, R.; Thaisrivongs, S.; Wlodawer, A. Crystal structure of a complex of HIV-1 protease with a dihydroxyethylene-containing inhibitor: Comparisons with molecular modeling. *Protein Sci.* **1992**, *1*, 1061–1072.

(30) Appelt, K. 1992, personal communication.

(31) Thompson, W.; Fitzgerald, P.; Holloway, K.; Emilio, E.; Darke, P.; McKeever, B.; Schleif, W.; Quintero, J.; Zugay, J.; Tucker, T.; Schwering, J.; Homnick, C.; Nunberg, J.; Springer, J.; Huff, J. Synthesis and antiviral activity of a series of HIV-1 protease inhibitors with functionality tethered to the P1 or P1′ phenyl substituents: X-ray crystal structure assisted design. *J. Med. Chem.* **1992**, *35*, 1685–1701.

(32) Fitzgerald, P.; McKeever, B.; Van Middlesworth, J.; Springer, J.; Dixon, R.; Darke, P. Crystallographic analysis of a complex between HIV-1 protease and acetyl-pepstatin at 2.0 angstroms resolution. *J. Biol. Chem.* **1990**, *265*, 14209–14219.

(33) Miller, M.; Schneider, J.; Sathyanarayana, B.; Toth, M.; Marshall, G.; Clawson, L.; Selk, L.; Kent, B.; Wlodawer, A. Structure of Complex of Synthetic HIV-1 Protease with a Substrate-Based Inhibitor at 2.3 Å Resolution. *Science* **1989**, *246*, 1149–1152.

(34) Jaskolski, M.; Tomasselli, A.; Sawyer, T.; Staples, D.; Heinrikson, R.; Schneider, J.; Kent, S.; Wlodawer, A. Structure at 2.5 Å resolution of chemically synthesized human immunodeficiency virus type-1 protease complexed with a human hydroxy-ethylene based inhibitor. *Biochemistry* **1991**, *30*, 1600–1609.

(35) Mimoto, T.; Imai, J.; Tanaka, S.; Hattori, N.; Takahashi, O.; Kisanuki, S.; Nagano, Y.; Shintani, M.; Hayashi, H.; Sakikawa, H.; Akaji, K.; Kiso, Y. Rational design and synthesis of a novel class of active site-targeted HIV protease inhibitors containing a hydroxymethylcarbonyl isostere. Use of phenylnorstatine or allophenylnorstatine as a transition-state mimic. *Chem. Pharm. Bull.* **1991**, *39*, 2465–2467.

(36) Cheng, Y.-c.; Prusoff, W. Relationship between the inhibition constant (Ki) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **1973**, *22*, 3099–3108.

(37) Tomasselli, A.; Olsen, M.; Hui, J.; Staples, D.; Sawyer, T.; Heinrikson, R.; Tomich, C.-S. Substrate analogue inhibition and active site titration of purified recombinant HIV-1 protease. *Biochemistry* **1990**, *29*, 264–269.

(38) Sawyer, T.; Tomasselli, A.; Poorman, R.; Hui, J.; Hinzmann, J.; Staples, D.; Maggiora, L.; Smith, C.; Heinrikson, R. In *Peptides - Chemistry, Structure and Biology*; Rivier, J., Marshall, G., Eds.; ESCOM: Leiden, 1989; pp 855–857.

(39) Wlodawer, A.; Erickson, J. Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* **1993**, *62*, 543–585.

(40) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.

(41) Oprea, T.; Cruciani, G.; Riganelli, D.; Clementi, S.; Marshall, G. Unpublished results.

(42) Mager, H.; Mager, P. Validation of QSARs: some reflections. *Quant. Struct.-Act. Relat.* **1992**, *11*, 518–521.

(43) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191–193.

(44) Topliss, J.; Edwards, R. Chance factors in studies of QSAR. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(45) Clark, M.; Cramer, R. The probability of chance correlation using PLS. *Quant. Struct.-Act. Relat.* **1993**, *12*, 137–145.

(46) Folkers, G.; Merz, A.; Rognan, D. CoMFA: scope and limitations. In *3D-QSAR in Drug Design*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 583–618.