# Enzyme Binding Selectivity Prediction: α-Thrombin vs Trypsin Inhibition[†]

G. Mlinsek,[‡] M. Novic,[§] M. Kotnik,[||] and T. Solmajer*,[‡,||]

Laboratories of Molecular Modeling and NMR Spectroscopy and of Chemometrics, National Institute of Chemistry, Hajdrihova 19, P.O. Box 660, 1001 Ljubljana, and Drug Discovery, Lek Pharmaceuticals d.d., Verovskova 57, 1526 Ljubljana, Slovenia

Received April 16, 2004

In the present work we explore the possibility of an in-depth computational analysis of available experimental X-ray structures in the specific case of a series of α-thrombin and trypsin complexes with their respective inhibitors for the development of a novel scoring function based on molecular electrostatic potential computed at the contact surface in the enzyme−inhibitor molecular complex. We subsequently employ the chemometrical approach to determine which are the interactions in the large volume of data that determine the resulting experimental binding constant between ligand and receptor. The results of the model evaluated with molecules in the independent validation set show that a reasonable average error of 1.30 log units of the difference between experimental and calculated binding constants was achieved in the system thrombin−trypsin, which is comparable with those of methods from the literature. Furthermore, by a careful preparation of the Kohonen top layer in the artificial neural network approach that is normally perceived as a “black box device”, we have been able to follow the implications of the structure of the inhibitor−enzyme complex for the inhibitor’s binding constant. The method appears to be suitable for evaluation of selectivity in structurally similar enzymatic systems, which is currently an important problem in drug design.

## 1. INTRODUCTION

Structure-based drug design has become an increasingly used approach for the discovery of novel bioactive molecules[1] and evolved to a valuable addition to the classical arsenal of medicinal chemistry. The crystal structures of the ligand−macromolecular receptor complex provide the experimental basis for numerous fruitful applications of this method that have already resulted in several drugs on the market. The estimation of the binding constant of virtual enzyme inhibitors from experimental structures of analogous compounds available in the structural database[2] enables directing the development of novel compounds with favorable binding properties, thus importantly facilitating the process of rational drug design. Central to successful prediction of a binding constant and forming the basis of in silico screening of a virtual set of molecules is an accurate estimation of the scoring function.[3,4]

Due to the importance of this concept, there have been numerous attempts at defining a scoring function described in the literature which can roughly be divided into three groups: functions based on fundamental principles of statistical mechanics,[5] scoring functions which simulate the physicochemical process of ligand binding to the protein surface and utilize empirical potential functions,[6−9] and methods which are based on QSAR principles but employ molecular fields for evaluation of the interaction between a ligand and its receptor.[10−12] Empirical potential functions form a foundation for methods which simulate the thermodynamic cycle of the formation of the ligand−protein complex and have been used in both free energy perturbation methodology and the thermodynamic integration approach.[13] Both approaches have demonstrated good accuracy in a priori evaluation of the binding constant; however, due to the necessity to evaluate a large number of ensemble energy states and large consumption of computational resources, they appear to be less practical for evaluation of a large number of molecules. A variation of this methodology called the master equation approach in which the free energy is decomposed into the sum of atom−atom interactions has been proposed.[14] Due to the fact that while free energy is a function of the thermodynamic state but its components such as solvation free energy and change of internal rotations and translations upon binding are not, this approach has been at the heart of considerable controversy.[15,16] Recently, empirical scoring functions have enjoyed a lot attention of researchers in the field in particular due to their speed and have met with considerable success in virtual screening of large molecular databases.[17−20] They have been developed by introduction of empirical weight factors for each of the free energy components obtained by fitting the experimental structures of ligand−protein complexes to the ligand’s binding affinity by use of multivariate regression methods. A somewhat weaker point of this approach stems from the fact that it is not clear to what extent such weight factors could be applied to the compounds belonging to classes structurally different from those of compounds used in the development of the weights.

* Corresponding author phone: +386-01-4760-277; fax: +386-01-4760-300; e-mail: tom.solmajer@ki.si.

[†] Abbreviations: QSAR, quantitative structure−activity relationship; QM/MM, quantum mechanics/molecular mechanics; ANN, artificial neural network; MEP, molecular electrostatic potential; CP, counterpropagation; CV-LOO, cross validation leave one out; GA, genetic algorithm; PRESS, predicted residual errors sum of squares; RMS, root mean square.

[‡] Laboratory of Molecular Modeling and NMR Spectroscopy, National Institute of Chemistry.

[§] Laboratory of Chemometrics, National Institute of Chemistry.

[||] Lek Pharmaceuticals d.d.

10.1021/ci0401017 CCC: $27.50
Published on Web 09/01/2004

Scoring functions based on principles of statistical mechanics such as potential of the mean force have been developed from structural data and use the atom–atom decomposition of the binding free energy.[21] Implicit in this approach is the assumption that the experimental structure of the ligand–protein complex is a global minimum of free energy and in a state of thermodynamic equilibrium. In such a case the large number of pair distribution functions of ligand–protein atoms enables statistical evaluation of the atom–atom pair distance distribution and computation of the resulting binding affinity by parametrization of each atom–atom pair.

3D QSAR approaches using molecular fields such as COMFA have correlated the binding affinity of a series of molecules with their electrostatic, steric, and hydrophobic properties by the evaluation of the interaction energy of a given molecule with a variety of reference fragments in space around a given compound.[10,11] The recent development AfMoC introduced the adaptation of the interaction potential to a specific protein active site and significantly enhanced the predictive power of the method.[12]

We set out to test our hypothesis that introduction of chemometric methodology in selection of crucial molecular interactions could be instrumental as the missing link in the puzzle of correlating the biological activity of a molecule with the three-dimensional structure of the ligand–receptor complex.[22–24] We assumed that the probability to describe accurately such complicated processes as intermolecular recognition would increase if we consider the use of a spatially defined molecular descriptor based on physical quantum principles such as the molecular electrostatic potential computed at the contact surface in the enzyme–inhibitor molecular complex and subsequently employ the chemometrical approach to determine which are the interactions in this large volume of data that determine the resulting experimental binding constant between the ligand and receptor. It was found that by a combined use of QM/MM methodology for estimation of the enthalpy contribution and additional descriptors such as log *P*, hydrophobic effects related to loss of entropy and descriptors relating to reduction of translational and rotatational freedom of molecules upon binding the best correlations of the atomic structure of the complex and the experimental value of the binding constant could be found. An artificial neural network and a genetic algorithm were both used to choose the pattern of MEP values, which produce the best correlation with the binding constant, i.e., the QSAR model. The interpretation of the resulting model appears to be greatly facilitated by the underlying physical principles since one can trace the interaction pattern between the inhibitor and enzyme onto the experimentally observed three-dimensional structure of the complex in contrast to empirical scoring functions based on pairwise atomic interactions, which make it difficult to see the structural implications of the cumulative function value.

In the present work we explore the possibility of an in-depth computational analysis of available experimental X-ray structures in the specific case of a series of thrombin and trypsin complexes with their respective inhibitors for the development of such a principle.

Thrombin is a multifunctional serine protease with trypsin-like specificity, and plays a central role in thrombosis and haemostasis by regulating the blood coagulation cascade and platelet activation processes.[25] Serving as the terminal enzyme of the cascade, thrombin cleaves fibrinogen to fibrin, which ultimately combines with platelets and other components to form a blood clot. The limited efficacy and the side effects of established antithrombotics, including warfarin, heparin, and acetylsalicylic acid, have provided the impetus for developing alternative anticoagulants.[26] Inhibition of thrombin is a prime target for therapeutic intervention of thrombosis. During the past decade, the search for orally bioavailable, potent, and selective low molecular weight thrombin inhibitors has become one of the most intensively studied areas in drug discovery.[27–28]

Substrate specificity of serine proteases is defined by the atomic structure of precleavage sites S1–S4 and primed pockets S1′–S4′after the hydrolyzed peptide bond.[29] Crystal structures of α-thrombin and trypsin have a considerable number of similarities but also some differences, which mediate the selectivity and are crucial in the design of selective thrombin inhibitors. On the basis of the classic tripeptide sequence D-Phe-Pro-Arg, which anchors at the S1–S3 pockets in a reverse manner, numerous thrombin inhibitors have been prepared. In particular, the details of the S1 binding site structure have been useful in the optimization strategy, which is frequently based on arginine substitution by guanidine and amidine mimetics and heteroaromatic arginine mimetics. The latter have been shown to possess interesting selectivity toward trypsin due to specific differences in the three-dimensional structure of the S1 pocket.

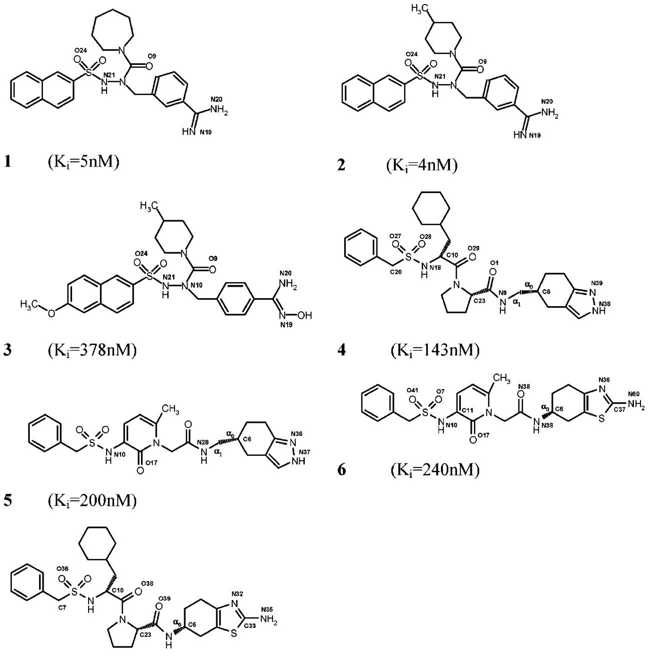
To utilize the availability of a large number of experimental structures of both α-thrombin and trypsin complexes with their respective inhibitors, we extended our previous work[24] to study the possibility of selectivity prediction and validate the approach on a larger dataset. Our conclusions substantiate the assumption that careful use of the chemometrics arsenal of data obtained by a straightforward application of QM/MM to experimental crystal structures of α-thrombin and trypsin with noncovalently bound inhibitors can result in prediction of the inhibitor selectivity of these structurally closely related molecular systems.

## 2. METHODS

The modules of a modular algorithm described in detail in our previous paper[24] are construction of a database of crystal structures of enzyme–inhibitor complexes, computation of MEP values on the boundary surface of inhibitor–enzyme complexes by using the QM/MM approach, reduction of the number of computed MEP values, CP-ANN training and testing, reduction of the structural representation by the use of a genetic algorithm, and independent validation of the model.

**a. Construction of the Database.** A complete search of the Brookhaven PDB for α-thrombin and trypsin structures was performed, and crystal structures of reversible noncovalent inhibitors were selected for further elaboration. Inhibitors which were found to bind covalently were not considered to exclude the possibility of intrinsic differences in the underlying biochemical mechanism. This process resulted in a list of 36 complexes of α-thrombin[33–54] and 32 structures of trypsin[55–67] with their respective inhibitors. In addition, molecules **1**–**7** (Chart 1) for which we have

**Chart 1.** Structures of Molecules **1**−**7** That Were Determined in Complex with α-Thrombin and Added into the Database[35,36,44,45]



**1** ($K_i$=5nM)

**2** ($K_i$=4nM)

**3** ($K_i$=378nM)

**4** ($K_i$=143nM)

**5** ($K_i$=200nM)

**6** ($K_i$=240nM)

**7** ($K_i$=128nM)

determined their crystal structure in complex with α-thrombin[35,36,44,45] were added to the database.[68]

The inhibitors in the series were not from a homologous structural series but composed from a variety of one to three building blocks (molecular fragments). All crystal structures in the database were carefully visually inspected and ligand modes of binding verified to bind into the respective enzymatic active site. The chemical structures of the inhibitors were quite heterogeneous in terms of size with between 21 and 118 atoms for thrombin and 18 to 145 atoms for trypsin inhibitors.

(i) To validate the role of different crystallization procedures used in various laboratories obtaining the experimental data and (ii) to estimate the influence of inhibitor binding on the geometry of the active site, we initially superimposed all enzyme structures onto a reference structure (structure 1A2C[53] was arbitrarily chosen as a reference). RMS values obtained for all heavy atoms in the superposition list (85 residues and 686 heavy atoms in total) ranged from 0.51 to 0.77 Å. Thus, on the basis of relatively small average RMS changes introduced by inhibitor binding, we concluded that binding of reversible inhibitors to the active site does not significantly perturb its geometry.

**b. MEP Computation.** The van der Waals surfaces of all atoms on the active site surface of the protein that were less than 10 Å away from any atom of any of the inhibitors were selected to form the enzymatic contact surface and about 30−40 points per atom on this surface randomly chosen.[24] Such point density was found to be sufficient for the smooth representation of MEP values for each atom in the contact surface. In these coordinate points on the enzyme surface MEP was calculated by using a standard QM/MM method.[22−24] The resulting MEP vector for each inhibitor has the same length between 11.000 and 12.000 coordinate points and the contact surface for each thrombin−inhibitor complex is thus represented by a reasonably large set of points in which the molecular electrostatic potential was computed. All inhibitor atoms were treated ab initio with the standard 6-31G basis set, which had been previously shown to be sufficiently accurate for calculation of MEP values.[24] The enzyme's electrostatic environment was treated by including the atomic charges of the enzyme atoms in the one-electron Hamiltonian of the complex for the Hartree−Fock−Roothan iterative computation of the inhibitor wave function/electron density.

**c. Construction of a Uniform Length Vector of the Computed MEP Values.** To reduce the number of input parameters for the artificial neural network to a practical size without suppressing the accuracy and retain a valid description of the computed electrostatic contact surface, the size of the MEP vector was reduced as described previously.[24] A uniform length vector of representative MEP values was chosen by an automatic procedure, with each atom at the contact surface being represented by two points and resulting in 597 points at the contact surface of α-thrombin and 660 points at the contact surface of trypsin.

The method in its present form appears rather robust since it operates on sufficiently small portions of the molecular surface where the variation in MEP values is not significant. Also, the MEP values for each inhibitor independently of size are stored in a vector of uniform length, which is a prerequsite for a simple application of chemometrics machinery.

**d. CP-ANN Training and Testing.** The MEP values served as inputs for a counterpropagation artificial neural network[31] (CP-ANN). In the first step of the ANN training the CV-LOO (cross validation leave one out) procedure was performed $n$ times by using $n - 1$ inhibitors in the database for prediction of the omitted inhibitor binding constant. The resulting distribution of binding constant predictions was used to place inhibitors into three groups: (i) a training set, (ii) a test set, and (iii) a validation set which was used as an independent external validation set. Such a validation set of compounds, which are removed from the database at the beginning of the training procedure, provides a measure of the quality of the final model. In addition, it can be used for an independent estimate of convergence properties of parameters which determine the CP-ANN model such as the number of neurons in the artificial neural network, maximal and minimal learning speed, number of training epochs, etc.

The training set of molecules was determined in such a way that it covers most of the information space of ligand−protein interactions. By using the test set, the parameters of the CP-ANN models and genetic algorithm are optimized.

The CP-ANN model was developed using the available database of 36 thrombin and 32 trypsin complexes. Accordingly, the network size was chosen as $n_x = 6$, $n_y = 6$. At the start of the ANN training cycle the first vector of MEP values representing the inhibitor was mapped to its position in the Kohonen layer randomly, and each succeeding set of MEP values was mapped to the position with most similarity. The similarity of two sets of MEP values was evaluated according to eq 1,

$$d_j = \sum_{i=1}^{m} (x_{si} - w_{ij})^2 \qquad (1)$$

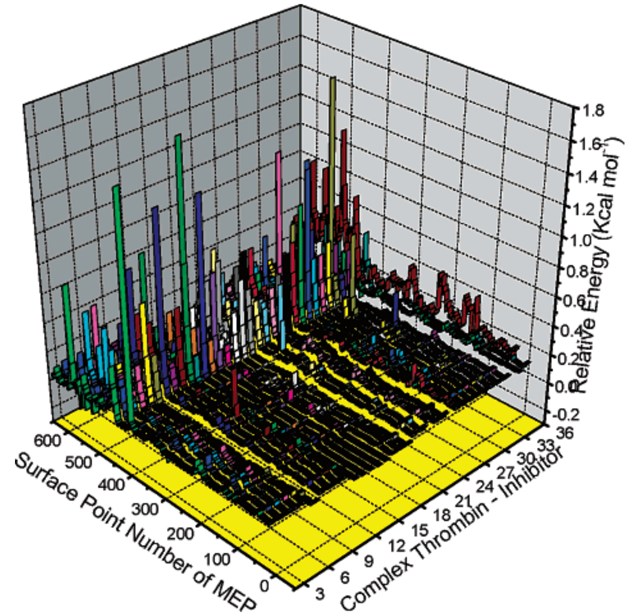


**Figure 1.** Values of MEP, computed with the QM/MM method in 594 points, for 36 thrombin inhibitors: *x* axis, surface point number in which MEP was computed; *y* axis, relative energy (kcal/mol); *z* axis, number of thrombin−inhibitor complexes (Table 1). The highest values are observed for amino acids Asp 189, Ala 190, Gly 216, and Gly 219, which delimit the S1 binding pocket of the active site.

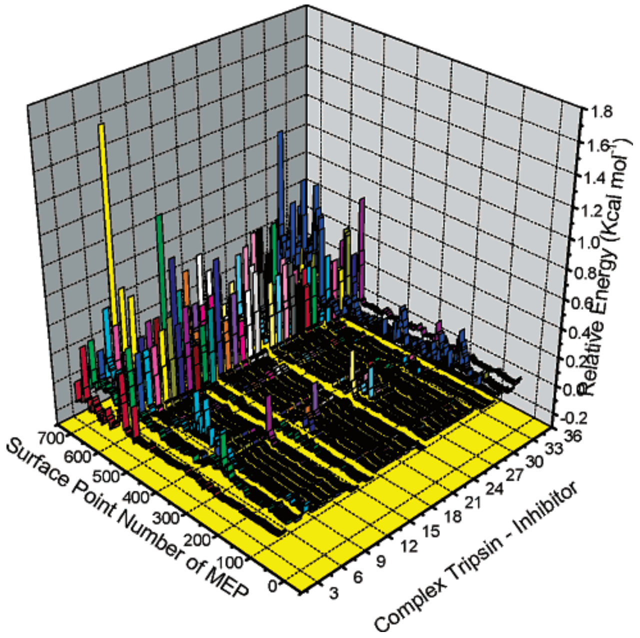


**Figure 2.** Values of MEP, computed with the QM/MM method in 660 points, for 32 trypsin inhibitors: *x* axis, surface point number in which MEP was computed; *y* axis, relative energy (kcal/mol); *z* axis, number of trypsin−inhibitor complexes (Table 2).

where $w_{ij}$ is the weight in the Kohonen layer, $x_{si}$ is the value of the MEP at position *s*, and *m* is the number of weights for a neuron. The weights in subsequent training steps are adapted according to eq 2,

$$w_{ij}^{(\mathrm{new})} = w_{ij}^{(\mathrm{old})} + \eta(t)\, a(d_c - d_j)\, [x_i - w_{ij}^{(\mathrm{old})}] \qquad (2)$$

where $\eta(t)$ determines the rate of learning [it is maximal at the beginning ($t = 1$, $\eta = a_{\max}$) and minimal at the end ($t = t_{\max}$, $\eta = a_{\min}$) of the learning procedure] and $a(d_c - d_j)$ is a function of the topological distance $d_c - d_j$ between the winning neuron *c* and input neuron *j*. The training cycle was repeated 125 times to ensure the convergence was reached.

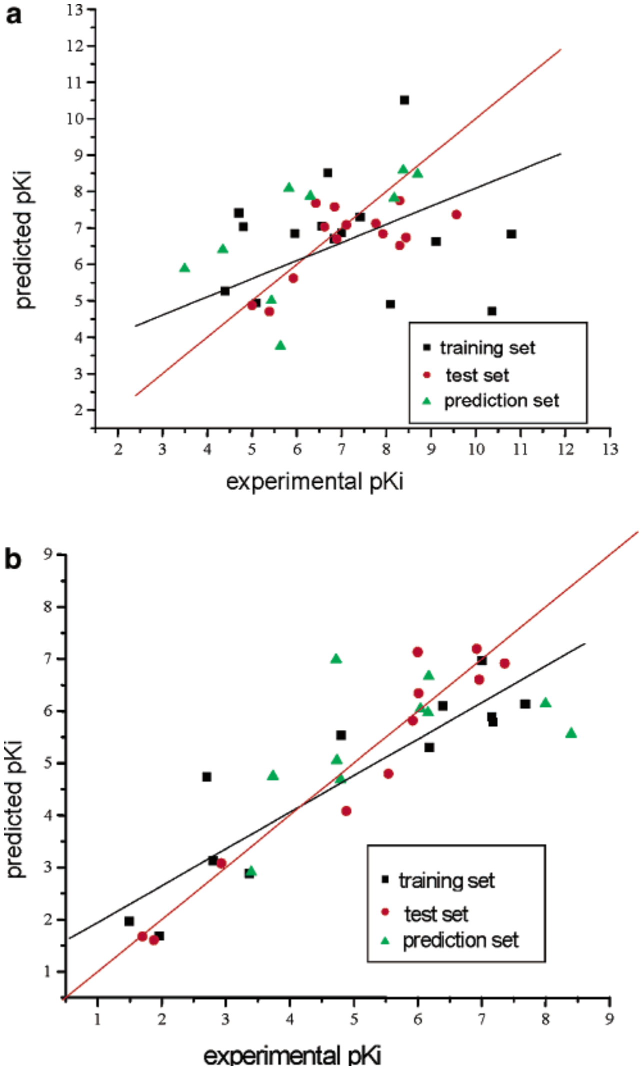
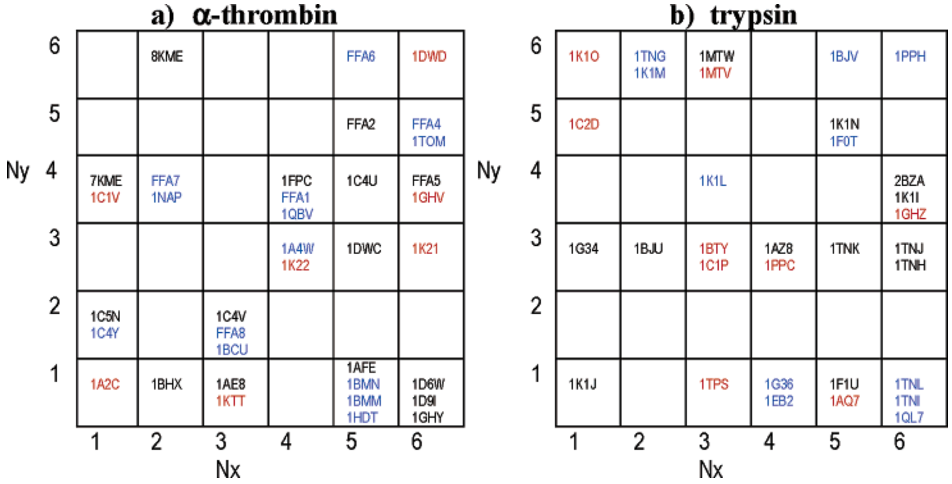


**Figure 3.** Results of the CV-LOO analysis: (a, top) thrombin inhibitors, (b, bottom) trypsin inhibitors. Experimental and computed values of the binding constant are given. On the basis of these data, the molecules were divided into training (■) and test (●) sets.

In the ANN testing procedure the difference between the target *T* and the predicted property *P* is squared and summed over all objects of the test set to give the PRESS (predicted residual errors sum of squares) measure of error. In addition, regression analysis yields the correlation coefficient *R* as an estimate of the quality of the derived model.

**e. Reduction of the Structural Representation by the Use of a Genetic Algorithm and Independent Validation of the Model.** The genetic algorithm procedure was used to reduce the complete contact surface for interaction of all inhibitors in the dataset with the enzyme to its most relevant part. A genetic algorithm mimicks three basic processes of Darwinian evolution: crossover, mutation, and selection. In the crossover step, new chromosomes are generated by mixing those of the parents. The mutation step describes random exchange of individual chromosome bits, and the



a) α-thrombin

| Ny \ Nx | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 6 | | 8KME | | | FFA6 | 1DWD |
| 5 | | | | | FFA2 | FFA4<br>1TOM |
| 4 | 7KME<br>1C1V | FFA7<br>1NAP | | 1FPC<br>FFA1<br>1QBV | 1C4U | FFA5<br>1GHV |
| 3 | | | | 1A4W<br>1K22 | 1DWC | 1K21 |
| 2 | 1C5N<br>1C4Y | | 1C4V<br>FFA8<br>1BCU | | | |
| 1 | 1A2C | 1BHX | 1AE8<br>1KTT | | 1AFE<br>1BMN<br>1BMM<br>1HDT | 1D6W<br>1D9I<br>1GHY |

b) trypsin

| Ny \ Nx | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 6 | 1K1O | 1TNG<br>1K1M | 1MTW<br>1MTV | | 1BJV | 1PPH |
| 5 | 1C2D | | | | 1K1N<br>1F0T | |
| 4 | | | 1K1L | | | 2BZA<br>1K1I<br>1GHZ |
| 3 | 1G34 | 1BJU | 1BTY<br>1C1P | 1AZ8<br>1PPC | 1TNK | 1TNJ<br>1TNH |
| 2 | | | | | | |
| 1 | 1K1J | | 1TPS | 1G36<br>1EB2 | 1F1U<br>1AQ7 | 1TNL<br>1TNI<br>1QL7 |

**Figure 4.** Map representation of the dataset with 36 fields ($N_x = 6$, $N_y = 6$) and distribution of molecules of the learning set (black), test set (blue), and validation set (red) in the map obtained by the CV-LOO model: (a) α-thrombin, (b) trypsin.

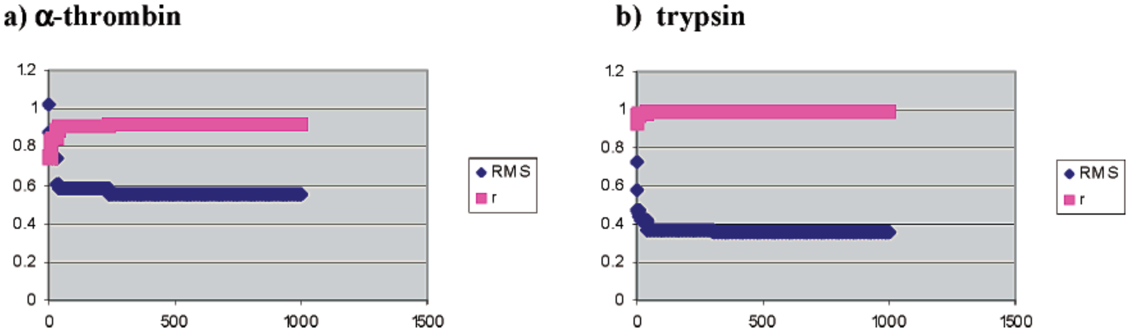


**Figure 5.** Representation of RMS value and the ascent of the correlation coefficient from a random starting point to the 900th generation in the genetic algorithm procedure: (a) α-thrombin, (b) trypsin. RMS convergence was achieved at the 237th generation for α-thrombin and at the 763th generation for trypsin.

selection step determines the chromosomes with optimal distribution of alleles for use in the next round. Each bit in the allele could take a value of 0 or 1. A value of 1 was used to assign the MEP value at a point in the model, while the value of 0 described the fact that the MEP value in a point on the contact surface was not considered.

The number of input parameters (597 for the inhibitors' electrostatic interaction with atoms of thrombin and 660 for interaction with trypsin) determines the length of the chromosome in bits. A pool of 100 chromosomes was tested using a random choice of the points on the surface. Seven alleles were given a value of 1 in a random fashion (for each experiment a new seed value was chosen). In the first generation for each chromosome a CP-ANN model was obtained. PRESS in the test set was used to determine the chromosome with the best predictive power. The best chromosomes were crossed over (mutated), and in the new pool of 100 chromosomes a new pattern was tested for quality. This procedure was repeated 900 times, and thus, in each generation a more representative pattern of contact points on the surface was obtained.

This last step should yield the lowest possible number of points on the enzymatic contact surface that gives a satisfactory prediction for $K_i$ for each individual inhibitor in the data set. In the resulting CP-ANN model the values of weights in the output layer enable the prediction of the inhibitor binding affinity.

## 3. RESULTS AND DISCUSSION

The MEP values for the α-thrombin–inhibitor and trypsin–inhibitor complexes were computed on the van der Waals surface of the same number of atoms (330 in the case of trypsin and 297 in thrombin) in each series. Since the database of available experimental structures was relatively small, we had to limit the representation of interaction at each atom on the surface to two points in the MEP vector. These values are plotted in Figures 1 and 2 for α-thrombin and trypsin inhibitors, respectively. A similar but in important details different pattern could be observed from Figure 2 in which the MEP values for the series of trypsin complexes are given.

High-energy values represent a significant electrostatic interaction between enzyme and inhibitor, while low values are a sign of hydrophobic contact or a large distance between the atomic pairs from which the interaction originates. By careful analyses of these graphs, it can be observed that the largest relative interaction energy is present at points on the contact surface which correspond to amino acids forming ionic and hydrogen bonds and which are in close proximity to the inhibitor. The highest MEP values in α-thrombin complexes are observed for atoms such as the carboxylic oxygen of Asp 189, carbonyl O of Ala 190 (Ser 190 in trypsin), and carbonyl oxygens of Gly 216 and Gly219. Significantly, these values are large for benzamidine inhibitors (0.8–0.9 kcal/mol), while at molecules with a neutral

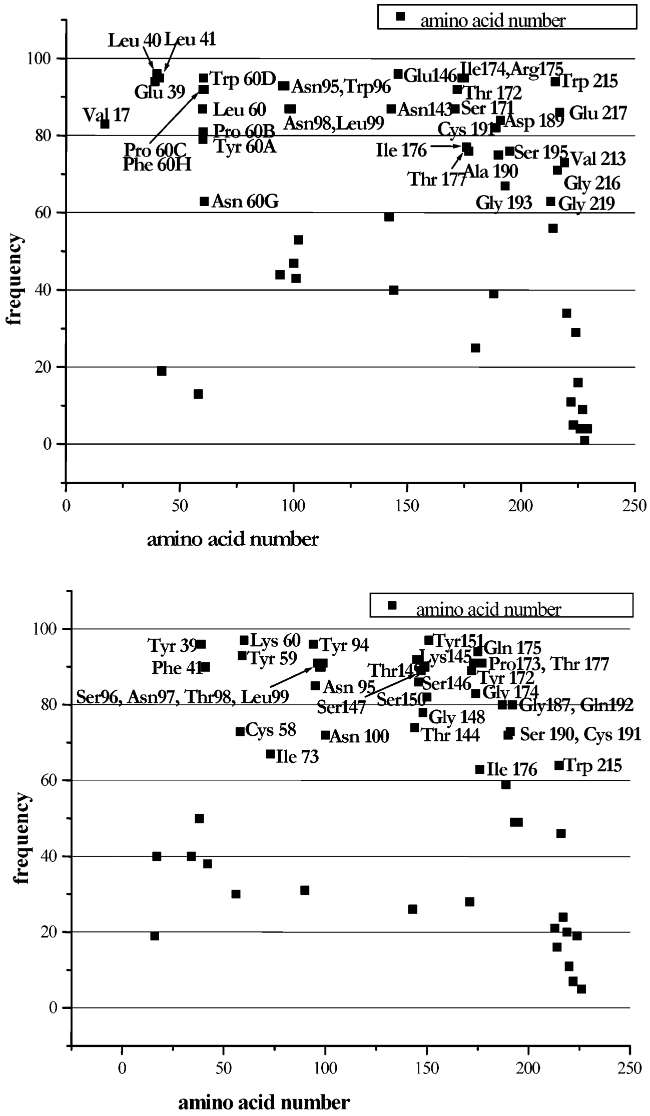


**Figure 6.** (a, top) Results of reduction of the MEP representation for thrombin by a genetic algorithm: frequency of the amino acid chosen by the algorithm out of all 54 amino acids in the active site to be important for correlation with the inhibitor binding constant. A total of 100 runs of GA were performed. (b, bottom) Same as (a) but for trypsin.

heterocycle at position S1 the corresponding MEP values are much smaller (0.1−0.2 kcal/mol).

Such qualitative observations between structure and MEP values can be made for other selected points on the contact surface in the inhibitor−enzyme complex as well. However, to simultaneously correlate such differences in MEP values of atoms for all amino acids in the active site of the inhibitor−enzyme complex and with the inhibitor's binding constant, we employed the combined ANN and GA approach.

The inhibitors were divided into three sets, training, test, and validation sets, to enable the independent determination of CP-ANN parameters (maximal and minimal speed of neural net learning, number of epochs, size of the net, correction of weights in neurons) from the first two sets and independent validation for the validation set.

The division into the training and test sets was based on CV-LOO results shown in parts a and b of Figure 3 for thrombin and trypsin inhibitors, respectively. The compounds which are poorly predicted in this scheme were included in the training set to account for uniqueness in the structure−property domain.[24] The ability of prediction expressed as $pK_{i,exptl} - pK_{i,calcd}$ was used as the criterion for placing the molecule in the training set, which should be able to cover most of the descriptor space. Furthermore, a uniform distribution of inhibitors in the network is important for its predictive ability and is shown in Figure 4. The third request was the maximum dispersion of $pK_i$ values that in our case spanned 7 orders of magnitude.

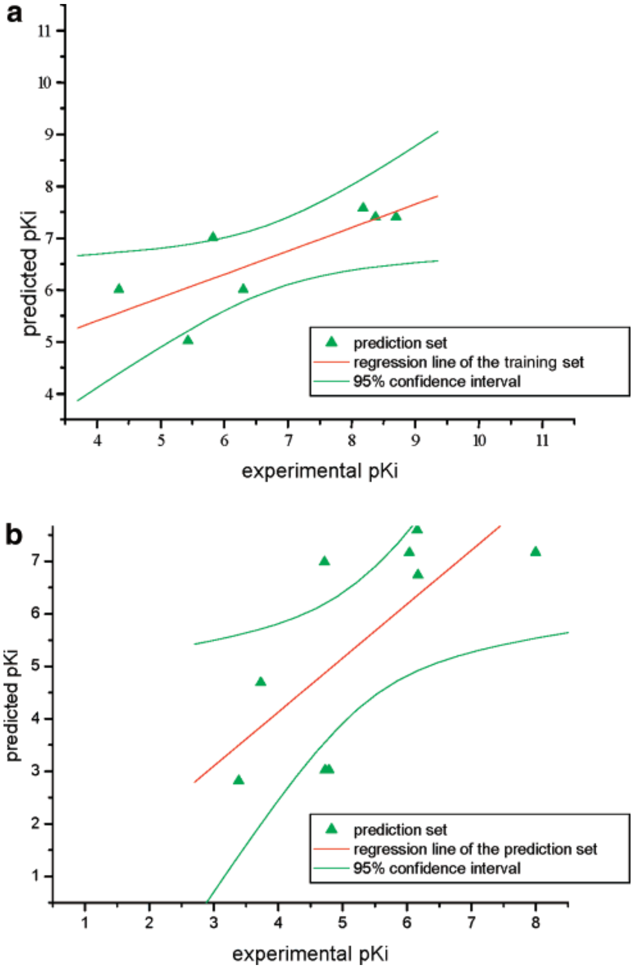


**Figure 7.** (a, top) CP-ANN model for prediction of $pK_i$ for thrombin inhibitors from a reduced representation (chosen chromosome) of MEP values with a 95% confidence limit for the validation set (seven inhibitors). (b, bottom) Same as (a) but for the trypsin validation set (nine inhibitors).

The final CP-ANN model was used for a search of the reduced representation of MEP values by the use of a genetic algorithm. Thus, from 54 amino acids forming the active site, those that play the most important role in binding were selected by using the genetic algorithm approach.

To achieve a useful statistical estimate, we performed 100 runs of searching for an optimal pattern of MEP representation. In each run, 7 vectors of MEP values were randomly seeded in the pool of 100 chromosomes and 900 generations of the genetic algorithm (mutations, crossover, and selection of optimal) were executed. A good correlation factor *R* and low values of RMS for the test set were achieved by using about 100 MEP vectors (98 for α-thrombin and 97 for trypsin). The convergence in correlation coefficient *R* and RMS value of the test set was achieved in the 237th

**Table 1.** Comparison of Experimental p$K_i$ Values for α-Thrombin with p$K_i$ Values Computed with the CP-ANN Model and p$K_i$ Values Computed with CP-ANN Models with Normalized Descriptors D: log *P*, Hydrophobic Area, Number of Rotatable Bonds[a]

| PDB code or compd no. | | p$K_{i,exptl}$ | p$K_{i,calcd}$(MEP) | Δp$K_{i,exptl-calcd}$(MEP) | p$K_{i,calcd}$(MEP + D) | Δp$K_{i,exptl-calcd}$(MEP+D) |
|---|---|---|---|---|---|---|
| 7KME | I | −4.39 | −4.88 | −0.48 | −6.08 | −1.68 |
| 1C5N | I | −4.69 | −4.89 | −0.19 | −5.94 | −1.24 |
| 1AFE | I | −4.79 | −4.90 | −0.11 | −4.89 | −0.09 |
| 8KME | I | −5.09 | −4.89 | 0.19 | −5.11 | −0.02 |
| 1D6W | I | −5.95 | −6.00 | −0.04 | −8.10 | −2.14 |
| 1AE8 | I | −6.55 | −6.61 | −0.06 | −6.53 | 0.01 |
| 5 | I | −6.69 | −6.65 | 0.04 | −6.56 | 0.13 |
| 1BHX | I | −6.83 | −7.57 | −0.73 | −7.65 | −0.81 |
| 1FPC | I | −7.00 | −6.98 | 0.01 | −6.19 | 0.80 |
| 1DWC | I | −7.40 | −8.48 | −1.07 | −8.47 | −1.07 |
| 1GHY | I | −8.09 | −8.09 | 0.00 | −5.94 | 2.15 |
| 2 | I | −8.39 | −8.32 | 0.07 | −8.40 | 0.00 |
| 1D9I | I | −9.10 | −9.92 | −0.81 | −9.05 | 0.05 |
| 1C4U | I | −10.36 | −10.24 | 0.11 | −8.10 | 2.25 |
| 1C4V | I | −10.79 | −9.92 | 0.86 | −10.73 | 0.06 |
| 1BCU | II | −5.00 | −5.01 | −0.01 | −5.94 | −0.94 |
| 1QBV | II | −5.38 | −4.88 | 0.50 | −6.19 | −0.80 |
| 1A4W | II | −5.92 | −5.97 | −0.05 | −5.92 | −0.00 |
| 3 | II | −6.42 | −6.65 | −0.23 | −6.56 | −0.13 |
| 6 | II | −6.61 | −6.65 | −0.03 | −6.61 | 0.00 |
| 4 | II | −6.84 | −6.65 | 0.18 | −6.86 | −0.02 |
| 7 | II | −6.89 | −6.91 | −0.02 | −6.86 | 0.02 |
| 1BMM | II | −7.10 | −7.11 | −0.01 | −7.13 | −0.02 |
| 1HDT | II | −7.76 | −7.75 | 0.00 | −6.08 | 1.68 |
| 1C4Y | II | −7.92 | −7.92 | −0.00 | −7.92 | −0.00 |
| 1 | II | −8.30 | −7.57 | 0.72 | −8.31 | −0.00 |
| 1TOM | II | −8.30 | −8.27 | 0.02 | −8.28 | 0.01 |
| 1BMN | II | −8.43 | −8.41 | 0.02 | −7.65 | 0.78 |
| 1NAP | II | −9.56 | −8.48 | 1.08 | −8.47 | 1.08 |
| 1GHV | III | −4.34 | −6.00 | −1.66 | −5.94 | −1.59 |
| 1C1V | III | −5.43 | −5.01 | 0.41 | −6.49 | −1.06 |
| 1KTT | III | −5.82 | −7.00 | −1.18 | −8.31 | −2.48 |
| 1A2C | III | −6.30 | −6.00 | 0.29 | −6.08 | 0.22 |
| 1DWD | III | −8.18 | −7.57 | 0.60 | −8.10 | 0.07 |
| 1K21 | III | −8.37 | −7.40 | 0.97 | −7.65 | 0.72 |
| 1K22 | III | −8.69 | −7.40 | 1.29 | −7.65 | 1.04 |

[a] Key: I, training set; II, test set; III, validation set.

generation for α-thrombin and 763rd generation for trypsin as shown in Figure 5.

The RMS and *R* values for the test set were used as an optimization criterion for the quality of the reduced representation, i.e., correlation between the experimental and calculated p$K_i$ values obtained by using the CP-ANN model for the molecules in the test set. The quality of the fit and correlation coefficient *R* converge toward 1 with each new epoch of training. These results were used for analyses of the frequency distribution of amino acids in the active site, which were chosen by the procedure to participate most frequently in determination of p$K_i$.

The amino acids chosen in more than 60% of the cases have been included in the model representation. The result is plausible since the amino acids present in the active site pockets S1−S4 are most frequent. Interestingly, the secondary sites S1′ and S2′ containing amino acids Glu39, Leu 40, and Leu 41 are represented with a frequency close to 90%. On the other hand, Asp 189 with a frequency of 84% is not completely present in the pattern, which can be explained by the presence of inhibitors in the test set that possess a neutral P1 functionality (charge 0) and thus are not capable of strong electrostatic interaction with site S1.

The above-described procedure resulted in the CP-ANN model, which served in evaluation of the predictive ability of $K_i$ for the validation set of α-thrombin and trypsin inhibitors. It is worthwhile to stress that for this purpose only the set of molecules serving in an independent assessment and completely excluded from the training procedure was taken into account.

The interactions in the inhibitor−enzyme complex as represented by the electrostatic descriptor MEP are a description of the enthalpy part of the binding process. To test also the significance of other factors for the complex formation such as solvation and change of entropy upon binding, we complemented the MEP by empirical descriptors such as log *P*, number of inhibitor free rotatable bonds, and hydrophobic surface of the inhibitors.[68] The results are summarized in Tables 1 and 2.

Analysis of the resulting Δp$K_i$ values shows that in the case of α-thrombin inhibitors the MEP model performs slightly better than the model in which MEP was supplemented with additional descriptors. In trypsin, the model MEP + D predicts the inhibitor binding constant more accurately. To illustrate this point in a more quantitative way, we present the statistical analysis for the training−test set and validation set in both enzyme systems in Table 3.

The correlation coefficient *R* for correlation of experimental and calculated p$K_i$ values is greater than 0.8 in all models, which shows a reasonable prediction ability for inhibitors possessing a low as well as high binding constant. In both models used (MEP and MEP with empirical

**Table 2.** Comparison of Experimental p$K_i$ Values for Trypsin with p$K_i$ Values Computed with the CP-ANN Model and p$K_i$ Values Computed with CP-ANN Models with Normalized Descriptors D: log $P$, Hydrophobic Area, Number of Rotatable Bonds[a]

| PDB code | | p$K_{i,exptl}$ | p$K_{i,calcd}$(MEP) | Δp$K_{i,exptl-calcd}$(MEP) | p$K_{i,calcd}$(MEP + D) | Δp$K_{i,exptl-calcd}$(MEP + D) |
|---|---|---|---|---|---|---|
| 1TNJ | I | −1.95 | −1.83 | 0.11 | −1.91 | 0.03 |
| 1MTW | I | −2.69 | −2.77 | −0.08 | −2.81 | −0.12 |
| 2BZA | I | −2.8 | −2.86 | −0.06 | −3.02 | −0.22 |
| 1TNH | I | −3.36 | −3.37 | −0.01 | −3.02 | 0.33 |
| 1BJU | I | −4.79 | −5.16 | −0.37 | −4.68 | 0.10 |
| 1TNK | I | −1.48 | −1.49 | −0.01 | −1.51 | −0.03 |
| 1K1I | I | −6.18 | −6.28 | −0.10 | −6.16 | 0.01 |
| 1K1N | I | −6.38 | −6.28 | 0.09 | −6.31 | 0.06 |
| 1AZ8 | I | −7 | −7.00 | 0.00 | −6.98 | 0.01 |
| 1F0U | I | −7.16 | −7.15 | 0.00 | −7.15 | 0.00 |
| 1G34 | I | −7.17 | −7.05 | 0.11 | −7.16 | 0.00 |
| 1K1J | I | −7.67 | −7.63 | 0.03 | −7.59 | 0.07 |
| 1TNI | II | −1.69 | −1.83 | −0.14 | −1.74 | −0.05 |
| 1TNL | II | −1.87 | −1.83 | 0.03 | −1.91 | −0.04 |
| 1TNG | II | −2.93 | −2.86 | 0.06 | −3.02 | −0.09 |
| 1QL7 | II | −4.88 | −4.91 | −0.03 | −4.92 | −0.04 |
| 1BJV | II | −5.53 | −5.16 | 0.36 | −5.53 | −0.00 |
| 1PPH | II | −5.92 | −5.91 | 0.00 | −5.85 | 0.06 |
| 1F0T | II | −6 | −6.01 | −0.01 | −6.03 | −0.03 |
| 1EB2 | II | −6 | −5.97 | 0.02 | −5.96 | 0.03 |
| 1K1L | II | −6.92 | −7.14 | −0.22 | −7.14 | −0.22 |
| 1G36 | II | −6.95 | −7.05 | −0.10 | −6.89 | 0.05 |
| 1K1M | II | −7.35 | −7.14 | 0.20 | −7.14 | 0.20 |
| 1MTV | III | −3.39 | −2.77 | 0.61 | −2.81 | 0.57 |
| 1C1P | III | −3.73 | −5.52 | −1.79 | −4.68 | −0.95 |
| 1C2D | III | −4.72 | −7.00 | −2.28 | −6.98 | −2.26 |
| 1BTY | III | −4.73 | −5.52 | −0.79 | −3.02 | 1.70 |
| 1GHZ | III | −4.79 | −5.52 | −0.73 | −3.02 | 1.76 |
| 1AQ7 | III | −6.04 | −5.96 | 0.07 | −7.15 | −1.11 |
| 1PPC | III | −6.16 | −7.63 | −1.47 | −7.59 | −1.43 |
| 1K1O | III | −6.17 | −4.91 | 1.25 | −6.73 | −0.56 |
| 1TPS | III | −8 | −7.63 | 0.36 | −7.16 | 0.83 |

[a] See footnote $a$ in Table 1.

**Table 3.** Results of the CP-ANN Model in Prediction of Inhibitor Binding to α-Thrombin and Trypsin[a]

| | descriptor | R | RMS |
|---|---|---|---|
| | α-Thrombin | | |
| training−test | MEP | 0.936 | 0.437 |
| validation | MEP | 0.784 | 1.029 |
| training−test | MEP + D | 0.802 | 0.973 |
| validation | MEP + D | 0.637 | 1.284 |
| | Trypsin | | |
| training−test | MEP | 0.998 | 0.142 |
| validation | MEP | 0.657 | 1.240 |
| training−test | MEP + D | 0.998 | 0.118 |
| validation | MEP + D | 0.716 | 1.363 |

[a] Correlation coefficients and RMS values for the training−test set and validation set are given. Descriptors D are as in Tables 1 and 2.

descriptors) the values of the correlation coefficient for the training−test set are considerably higher than the values for the validation set, as expected. Interestingly, the correlation coefficient in α-thrombin was higher with the use of the MEP representation ($r = 0.78$), while for trypsin better correlation ($r = 0.71$) was obtained by using MEP with additional descriptors which take into account solvation, hydrophobic effects, and loss of rotational entropy. The difference between experimental and calculated Δp$K_i$ values (|p$K_{i,calcd}$ − p$K_{i,exptl}$|) is measured by the RMS value. Regardless of the model used, the RMS values for the training−test set are lower than the values for the validation set. The differences for the training−test set are less than 1 log unit in p$K_i$ for all models considered, while for the validation set these values are slightly higher (the RMS value varies from 1.02 to 1.58 log units) but still reasonable.

To verify that the ANN is not overtrained, we added the test set to the training set, and the fact that the correlation coefficient $R$ of the training set did not increase and the $R$ of the test set did not decrease shows that the network is not overtrained.[69]

The results presented above show that the method CP-ANN with use of the GA for selection of important MEP values from the bulk pool of computed MEP values is an efficient method for extracting the correlation between experimental structure and binding constant. On the other hand, the comparison of correlation coefficients $R$ and RMS values for the training−test set and validation set points to a deficiency of the model which can be traced to the dependence of the model on the size of the database of structures and variability of structures present in the training−test set.

Analysis of computational approaches to free energy of binding in the literature displays a variety of views with respect to the importance of contributions to this complex phenomenon. Some authors defend the view that the importance of electrostatics and hydrogen bonding in particular is exaggerated in comparison with hydrophobic interactions and entropy.[71] The MEP as a descriptor of electrostatic interactions between a ligand and an enzyme is a relatively realistic representation of the physicochemical process of binding since in the area of predominantly hydrophobic interactions its value is close to 0, and these effects can be traced down in the MEP representation

**Table 4.** Comparison of Calculated Binding Constants of Inhibitors to Thrombin and Trypsin with the Results of Scoring Functions from FlexX

| PDB code | F-score | G-score | PMF | D-score | ChemScore | $pK_{i,calcd}$(MEP) | $pK_{i,calcd}$(MEP + D) | $pK_{i,exptl}$ |
|---|---|---|---|---|---|---|---|---|
| | | | | α-Thrombin | | | | |
| 1GHV | −30.14 | −138.52 | −40.93 | −93.50 | −31.00 | −6.00 | −5.94 | −4.34 |
| 1C1V | −20.43 | −169.40 | −65.35 | −118.12 | −21.65 | −5.01 | −6.49 | −5.43 |
| 1KTT | −38.32 | −290.41 | −52.82 | −169.20 | −33.94 | −7.00 | −8.31 | −5.82[a] |
| 1A2C | −3.40 | −305.33 | −56.38 | −186.73 | −13.73 | −6.00 | −6.08 | −6.30[a] |
| 1DWD | −33.13 | −273.76 | −74.41 | −175.97 | −34.04 | −7.57 | −8.10 | −8.18 |
| 1K21 | −22.52 | −273.74 | −81.47 | −156.57 | −25.23 | −7.40 | −7.65 | −8.37 |
| 1K22 | −38.94 | −312.35 | −70.51 | −169.00 | −29.86 | −7.40 | −7.65 | −8.69 |
| | | | | Trypsin | | | | |
| 1MTV | −28.03 | −211.76 | −19.41 | −135.18 | −28.46 | −2.77 | −2.81 | −3.39[a] |
| 1C1P | −27.30 | −138.54 | −29.20 | −84.20 | −23.91 | −5.52 | −4.68 | −3.73 |
| 1C2D | −16.36 | −144.50 | −33.44 | −83.27 | −21.52 | −7.00 | −6.98 | −4.72 |
| 1BTY | −23.83 | −109.52 | −2.51 | −64.90 | −23.85 | −5.52 | −3.02 | −4.73 |
| 1GHZ | −26.12 | −128.55 | −17.52 | −77.63 | −24.76 | −5.52 | −3.02 | −4.79 |
| 1AQ7 | −11.44 | −291.34 | −32.86 | −166.37 | −15.84 | −5.96 | −7.15 | −6.04[a] |
| 1PPC | −28.68 | −229.07 | −39.79 | −146.24 | −27.83 | −7.63 | −7.59 | −6.16 |
| 1K1O | −10.82 | −229.35 | −36.57 | −145.30 | −18.35 | −4.91 | −6.73 | −6.17 |
| 1TPS | −35.62 | −459.65 | −73.50 | −226.05 | −20.79 | −7.63 | −7.16 | −8.00[a] |

[a] log $IC_{50}$.

**Table 5.** Statistical Parameters for Correlation between Experimental and Calculated $pK_i$ Values for Prediction of Inhibitor Binding to Thrombin ($N = 7$) and Trypsin ($N = 9$), Obtained by the CP-ANN Model and by the Scoring Functions from the Program FlexX

| | F-score | G-score | PMF | D-score | ChemScore | $pK_{i,calcd}$(MEP) | $pK_{i,calcd}$(MEP + D) |
|---|---|---|---|---|---|---|---|
| | | | | α-Thrombin | | | |
| correlation coefficient $R$ | 0.25 | 0.73 | 0.82 | 0.26 | 0.62 | 0.78 | 0.63 |
| standard deviation $\sigma$ | | | | | | 1.10 | 1.30 |
| significance $p$ | 0.58 | 0.062 | 0.022 | 0.566 | 0.137 | 0.036 | 0.123 |
| | | | | Trypsin | | | |
| correlation coefficient $R$ | −0.49 | 0.81 | 0.75 | 0.77 | 0.59 | 0.67 | 0.71 |
| standard deviation $\sigma$ | | | | | | 1.19 | 1.43 |
| significance $p$ | 0.173 | 0.008 | 0.017 | 0.015 | 0.109 | 0.046 | 0.030 |

(Figures 1 and 2). However, the effect of entropy and solvation is not represented by this model, and we have additionally tested empirical descriptors log $P$, the loss of hydrophobic area of inhibitors upon binding, and the number of rotatable bonds for an estimate of the loss of rotational entropy for both enzymatic systems α-thrombin and trypsin. Our study enables the discussion on the influence of structure diversity of the ligands and incorporation of empirical descriptors in the model. For trypsin, the inhibitor structures were more structurally homogeneous, and thus, their MEP similarity in the training set was larger than in α-thrombin. This homogeneity is reflected in a better correlation coefficient in the training set ($R = 0.99$ (RMS = 0.11) for MEP alone and $R = 0.99$ (RMS = 0.14) for MEP with empirical descriptors). In the validation set, however, the correlation coefficient is worse ($R = 0.65$ (RMS = 1.24) for MEP alone and $R = 0.71$ (RMS = 1.36) for MEP with empirical descriptors).

In the α-thrombin training set the inhibitor structures are more diverse; thus, the resulting correlation is worse ($R = 0.93$ (RMS = 0.43) for MEP alone and $R = 0.84$ (RMS = 0.86) for MEP with addition of empirical descriptors). The prediction ability, however, compares favorably with that of trypsin since in the validation set we obtain $R = 0.78$ (RMS = 1.02) for the MEP model and $R = 0.63$ (RMS = 1.28) for MEP with additional descriptors.

We note that the variability of MEP values in the α-thrombin system (Figure 1) could partially be interpreted by the role of the YPPW loop in ligand binding. This hydrophobic loop, which is specific for α-thrombin among serine proteases, delimits the motion of the inhibitor in the S2 pocket. On the other hand, the S1 pocket of α-thrombin is larger than in trypsin since the Ser 190 side chain of trypsin is exchanged for Ala 190 in α-thrombin. In both enzymatic systems the additional empirical descriptors do not improve the prediction ability of the CP-ANN model, which could be explained by predominant role of electrostatics in inhibitor binding.

In conclusion we compare the results of the CP-ANN model with binding constants obtained by the use of the docking algorithm FlexX.[6−9,12] A variety of scoring functions are available, and in Table 4 we present the results for a set of available scoring functions used for inhibitors in the validation set and compare them with calculated $pK_i$ values through use of MEP and MEP with additional empirical descriptors.

The advantage of FlexX is in using several scoring functions, which can result in a better correlation of experimental and calculated $pK_i$ values for a selected scoring function. However, the resulting scores are relative values of an energy evaluation, calculated for each inhibitor, while the CP-ANN model yields absolute values of predicted binding constants $pK_i$. The statistical comparison of scores obtained with scoring functions as available in FlexX and predicted $pK_i$ values from the CP-ANN model presented in Table 5 shows that both approaches predict $pK_i$ with an expected standard deviation of 1−1.5 $pK_i$ units.

Most docking algorithms perform well in orienting the candidate molecule in the active site in the proper binding pose. However, the accurate determination of the binding energy is notoriously difficult, and the accuracy of the predictions varies greatly.[72]

## 4. CONCLUSIONS

On the basis of experimental structural data of inhibitor−enzyme complexes, a scoring function has been developed that enables prediction of the binding constant for virtual inhibitors. As a descriptor for the interaction between inhibitor and the protein, the molecular electrostatic potential at the enzymatic active site surface was used. The biological activities of molecules in the training set were correlated with MEP by using a model based on artificial neural networks and reduction of the number of MEP values by a genetic algorithm. The results of the model evaluated with molecules in the independent validation set show that a reasonable average error of 1.30 log units of the difference between experimental and calculated binding constants was achieved in the thrombin−trypsin system, which is comparable with those of methods from the literature. An important novelty and advantage of this approach compared with previously published methods is that an absolute value of p$K_i$ is obtained by our procedure. Furthermore, by a careful preparation of the Kohonen top layer in the artificial neural network approach that is normally perceived as a “black box device”, we have been able to follow the implications of the structure of the inhibitor−enzyme complex for the inhibitor’s binding constant. It is clear that inclusion of a still more diversified set of inhibitors in the training−test set with less similarity in chemical structure and binding affinity would increase the degree of optimization of neural network parameters and increase the predictive ability of p$K_i$ by the use of the CP-ANN model. The method appears to be suitable for evaluation of selectivity in structurally similar enzymatic systems, which is currently an important problem in drug design.

## ACKNOWLEDGMENT

The Ministry of Science and Technology of Slovenia is thanked for financial support to G.M. and M.K.

**Supporting Information Available:** List of 36 α-thrombin−inhibitor and 32 trypsin−inhibitor complex crystallographic structures used in this study with references and binding constants of the inhibitor to thrombin, $K_i$ (expressed in log units), log $P$ values, sums of the inhibitor hydrophobic area, HAI, losses of hydrophobic area upon binding, LHAI, numbers of free rotating bonds, and total charges of the inhibitors and figures illustrating the results of the CP-ANN model for the training−test set for α-thrombin and trypsin. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Gubernator, K.; Bohm, H. J. In *Structure-based Ligand Design*; Gubernator, K., Bohm, H. J., Eds.; Methods and Principles in Medicinal Chemistry, Vol. 6; Wiley-VCH: Weinheim, Germany, 1998; pp 1−11.

(2) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; et al. Protein Data Bank. Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(3) Lyne P. D. Structure-based virtual screening: an overview. *Drug. Discovery Today* **2002**, *7*, 1047−1055.

(4) Stahl, M.; Rayer, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(5) Ajay; Murcko M. A. Computational Methods to Predict Binding Free Energy in Ligand−Receptor Complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.

(6) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(7) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(8) Bohm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D data base search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.

(9) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(10) Cramer, R. D.; DePriest, S. A.; Patterson, D. E.; Hecht, P. In *The Developing Practice of Comparative Molecular Field Analysis. 3D QSAR in Drug Design Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993; pp 443−485.

(11) Klebe, G. Comparative molecular similarity indices analyiss: CoMSIA. Perspect. *Drug Discovery* **1998**, *12*, 87−104.

(12) Gohlke, H.; Klebe, G. DrugScore Meets CoMFA: Adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45*, 4153−4170.

(13) Beveridge, D. L.; Di Capua, F. M. In *Computer Simulations of Biomolecular Systems*; van Gunsteren, W. F., Weiner, P. K., Eds.; ESCOM: Leiden, The Netherlands, 1989; pp 1−26.

(14) Williams, D. H.; Cox, J. P. L.; Doig, A. J.; et al. Toward the semiquantitative estimation of binding constants. *J. Am. Chem. Soc.* **1991**, *113*, 7020−7030.

(15) Kollman, P. A.; Theory of Macromolecule-Ligand Interactions *Curr. Opin. Struct. Biol.* **1994**, *4*, 240−245.

(16) Mark, A. E.; van Gunsteren, W. F. Decomposition of the Free-Energy of a system interms of specific Interactions-Implications for Theoretical and Experimental Studies *J. Mol. Biol.* **1994**, *240*, 167−176.

(17) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based Scoring Function to Predict Protein−Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(18) Mitchell, J. B. O.; Laskowski, R. A.; Alexander, A.; Thornton, J. M. BLEEP—Potential of Mean Force Describing Protein-Ligand Interactions: I. Generating Potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(19) DeWitte, R. S.; Shakhnovich, E. I. SMoG: de Novo Design Method Based on Simple, Fast, and Accurate Free Eenergy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.

(20) Charifson, P. S.; Corkery, J. P.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(21) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(22) Bagdassarian, C. K.; Schramm, V. L.; Schwarty, S. D. Molecular electrostatic potential analysis for enzymatic substrates, competitive inhibitors and transition-state inhibitors. *J. Am. Chem. Soc.* **1996**, *118*, 8825−8836.

(23) Braunheim, B. B.; Miles, R. W.; Schramm, V. L.; Schwartz, S. D. Prediction of inhibitor binding free energies by quantum neural networks. Nucleoside Analogues binding to Trypanosomal Nucleoside Hydrolase. *Biochemistry* **1999**, *38*, 116076−16083.

(24) Mlinsek, G.; Novic, M.; Hodoscek, M.; Solmajer, T. Prediction of enzyme binding: human thrombin inhibition study by quantum chemical and artificial intelligence methods based on X-ray structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1286−1294.

(25) Rewinkel, J. B. M.; Adang, A. E. P. Strategies and progress towards the ideal orally active thrombin inhibitor. *Curr. Pharm. Des.* **1999**, *5*, 1043−1075.

(26) Lefkovits, J.; Topol, E. J. Direct thrombin inhibitors in cardiovascular medicine. *Circulation* **1994**, *3*, 1522−36.

(27) Stubbs, M.; Bode, W. A player of many parts: the spotlight falls on thrombin’s structure. *Thromb. Res.* **1993**, *69*, 1−58.

(28) Stone, S. R.; Tapparelli, C. Thrombin inhibitors as antithrombotic agents: the importance of rapid inhibition. *J. Enzyme Inhib.* **1995**, *9*, 3−15.

(29) Bode, W.; Mayr, I.; Baumann, U.; Huber, R.; Stone, S. R.; Hofsteenge, J. The refined 1.9 A crystal structure of human alpha-thrombin:

interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J*. **1989**, *8*, 3467−3475.

(30) *InsightII*, Version 97.0; MSI Inc.: San Diego, CA, 1997.

(31) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, Germany, 1999.

(32) Devilers, J. *Genetic Algorithms in Molecular Modeling*; Academic Press: New York, 1996.

(33) Mochalkin, I.; Tulinsky, A. Structure of thrombin retro-inhibited with SEL2711 and SEL2770 as they relate to factor Xa binding. *Acta Crystallogr., D* **1999**, *55*, 785−793.

(34) Banner, D. W.; Hadvary, P. Crystallographic Analysis at 3.0-A Resolution of the Binding to Human Thrombin of Four Active Site-directed Inhibitors. *J. Biol. Chem*. **1991**, *266*, 20085−20093.

(35) Zega, A.; Mlinšek, G.; Šolmajer, T.; Trampuš-Bakija, A.; Stegnar, M.; Urleb, U. Thrombin inhibitors built on an azaphenylalanine scaffold *Bioorg. Med. Chem. Lett*. **2004**, *14*, 1563−1567.

(36) Peterlin-Masic, L.; Mlinsek, G.; Solmajer, T.; Trampus-Bakija, A.; Stegnar, M.; Kikelj, D. Novel Thrombin Inhibitors Incorporating Nonbasic Partially Saturated Heterobicyclic P1-arginine mimetics. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 789−794.

(37) De Simone, G.; Balliano, G.; Milla, P.; et al. Human Alpha-Thrombin Inhibition by the Highly Selective Compounds N-ethoxycarbonyl-D-Phe-Pro-alpha-azaLys p-Nitrophenyl Ester and N-carbobenzoxy-Pro-alpha-azaLys p-Nitrophenyl Ester: A Kinetic, Thermodynamic and X-ray Crystallographic Study. *J. Mol. Biol*. **1997**, *269*, 558−569.

(38) Wagner, J.; Kallen, J.; Ehrhardt, C.; Evenou, J.-P.; Wagner, D. Rational Design, Synthesis, and X-ray Structure of Selective Noncovalent Thrombin Inhibitors. *J. Med. Chem*. **1998**, *41*, 3664−3674.

(39) Krishnan, R.; Mochalkin, I.; Arni, R.; Tulinsky A. Structure of thrombin complexed with selective non-electrophilic inhibitors having cyclohexyl moieties at P1. *Acta Crystallogr., D* **2000**, *56*, 294−303.

(40) Katz, B. A.; Mackman, R.; Luong, C.; et al. Structural Basis for Selectivity of a Small Molecule, S1-Binding, Submicromolar Inhibitor of Urokinase-Type Plasminogen Activator. *Chem. Biol*. **2000**, *7*, 299−306.

(41) Matthews, J. H.; Krishnan, R.; Costanzo, M. J.; Maryanoff, B. E.; Tulinsky, A. Crystal Structures of Thrombin with Thiazole-Containing Inhibitors: Probes of the S1′ Binding Site. *Biophys. J*. **1996**, *71*, 2830−2839.

(42) Malley, M. F.; Tabernero, L.; Chang, C. Y.; et al. Crystallographic determination of the structures of human alpha-thrombin complexed with BMS-186282 and BMS-189090. *Protein Sci*. **1996**, *5*, 221−228.

(43) Tabernero, L.; Chang, C. Y. Y.; Ohringer, S. L.; et al. Structure of a Retro-binding Peptide Inhibitor Complexed with Human Alpha-Thrombin. *J. Mol. Biol*. **1995**, *246*, 14−20.

(44) Peterlin-Masic, L.; Kranjc, A.; Mlinsek, G.; Solmajer, T.; Stegnar, M.; Kikelj D. Selective 3-Amino-2-Pyridinone Acetamide Thrombin Inhibitors Incorporating Weakly Basic Partially Saturated Heterobicyclic P1-Arginine Mimetics. *Bioorg. Med. Chem. Lett*. **2003**, *13*, 3171−3179.

(45) Marinko, P.; Krbavcic, A.; Mlinsek, G.; Šolmajer, T.; Trampuš Bakija, A.; Stegnar, M.; Stojan, J.; Kikelj. D. Novel non-covalent thrombin inhibitors incorporating P1 4,5,6,7-tetrahydobenzothiazole arginine mimetics. *Eur. J. Med. Chem*. **2004**, *39*, 257−265.

(46) Banner, D. W. Private communication.

(47) Conti, E.; Rivetti, C.; Wonacott, A.; Brick, P. X-ray and spectrophotometric studies of the binding of proflavin to the S1 specificity pocket of human alpha-thrombin. *FEBS Lett*. **1998**, *425*, 229−233.

(48) Bone, R.; Lu, T.; Illig, C. R.; Soll, R. M.; Spurlino, J. C. Structural Analysis of Thrombin Complexed with Potent Inhibitors Incorporating a Phenyl Group as a Peptide Mimetic and Aminopyridines as Guanidine Substitutes. *J. Med. Chem*. **1998**, *41*, 2068−2075.

(49) Lyle, T. A.; Chen, Z. G.; Appleby, S. D.; et al. Synthesis, evaluation, and crystallographic analysis of L-371,912: A potent and selective active-site thrombin inhibitor. *Bioorg. Med. Chem. Lett*. **1997**, *7*, 67−72.

(50) Dullweber, F.; Stubbs, M. T.; Musil, D.; Sturzebecher, J.; Klebe, G. Factorising Ligand Affinity: A Combined Thermodynamic and Crystallographic Study of Trypsin and Thrombin Inhibition. *J. Mol. Biol*. **2001**, *313*, 593−614.

(51) Katz, B. A.; Clark, J. M.; Finer-Moore, J. S.; et al. Design of potent selective zinc-mediated serine protease inhibitors. *Nature* **1998**, *391*, 608−612.

(52) Katz, B. A.; Elrod, K.; Luong, C.; et al. A Novel Serine Protease Inhibition Motif Involving a Multi-centered Short Hydrogen Bonding Network at the Active Site. *J. Mol. Biol*. **2001**, *307*, 1451−1486.

(53) Steiner, J. L. R.; Murakami, M.; Tulinsky A. Structure of Thrombin Inhibited by Aeruginosin 298-A from a Blue-Green Alga. *J. Am. Chem. Soc*. **1998**, *120*, 597−598.

(54) Hauel, N.; Nar, H.; Priepke, H.; Ries, U.; Stassen, J. M.; Wienen, W. Structure-Based Design of Novel Potent Nonpeptide Thrombin Inhibitors. *J. Med. Chem.* **2002**, *45*, 1757−1766.

(55) Stubbs, M. T.; Huber, R.; Bode W. Crystal structures of factor Xa specific inhibitors in complex with trypsin: structural grounds for inhibition of factor Xa and selectivity against thrombin. *FEBS Lett.* **1995**, *375*, 103−107.

(56) Maignan, S.; Guilloteau, J. P.; Pozieux, S.; et al. Crystal Structures of Human Factor Xa Complexed with Potent Inhibitors. *J. Med. Chem.* **2000**, *43*, 3226−3232.

(57) Presnell, S. R.; Patil, G. S.; Mura, C.; et al. Oxyanion-Mediated Inhibition of Serine Proteases. *Biochemistry* **1998**, *37*, 17068−17081.

(58) Maduskuie, T. P.; McNamara, K. J.; Ru, Y.; Knabb, R. M.; Stouten, P. F. W. Rational design and synthesis of novel, potent bisphenylamidine carboxylate factor Xa inhibitors *J. Med. Chem*. **1998**, *41*, 53−62.

(59) Kurinov, I. V.; Harrison, R. W. Prediction of new serine proteinase inhibitors. *Nat. Struct. Biol*. **1994**, *1*, 735−743.

(60) Ota, N.; Stroupe, C.; Ferreira-da-Silva, J. M. S.; Shah, S. A.; Mares-Guia M. Non-Boltzmann Thermodynamic Integration (NBTI) for Macromolecular Systems: Relative Free Energy of Binding of Trypsin to Benzamidine and Benzylamine. *Proteins: Struct., Funct., Genet*. **1999**, *37*, 641−653.

(61) Nar, H.; Bauer, M.; Schmid, A.; et al. Structural basis for Inhibition Promisuity of Dual Specific Thrombin and Factor Xa Blood Coagulation Inhibitors. *Structure* **2001**, *9*, 29.

(62) Turk, D.; Sturzebecher, J.; Bode W. Geometry of binding of the N alpha-tosylated piperidides of m-amidino-, p-amidino- and p-guanidino phenylalanine to thrombin and trypsin. X-ray crystal structures of their trypsin complexes and modeling of their thrombin complexes. *FEBS Lett*. **1991**, *287*, 133−138.

(63) Liebeschuetz, J. W.; Jones, S. D.; Morgan, P. J.; et al. PRO_SELECT: Combining Structure-Based Drug Design and Array-Based Chemistry for Rapid Lead Discovery. 2. The Development of a Series of Highly Potent and Selective Factor Xa Inhibitors. *J. Med. Chem*. **2002**, *45*, 1221−1232.

(64) Bode, W.; Turk, D.; Sturzebecher, J. Geometry of binding of the benzamidine- and arginine-based inhibitors N alpha-(2-naphthyl-sulphonyl-glycyl)-DL-p-amidinophenylalanyl-pipe ridine (NAPAP) and (2R,4R)-4-methyl-1-[N alpha-(3-methyl-1,2,3,4-tetrahydro-8- quinolinesulphonyl)-L-arginyl]-2-piperidine carboxylic acid (MQPA) to human alpha-thrombin. X-ray crystallographic determination of the NAPAP-trypsin complex and modeling of NAPAP-thrombin and MQPA-thrombin. *Eur. J. Biochem*. **1990**, *193*, 175−182.

(65) Katz, B. A.; Finer-Moore, J.; Mortezaei, R.; Rich, D. H.; Stroud, R. M. Episelection: Novel Ki Nanomolar Inhibitors of Serine Proteases Selected by Binding or Chemistry on an Enzyme Surface. *Biochemistry* **1995**, *34*, 8264−8280.

(66) Lee, A. Y.; Smitka, T. A.; Bonjouklian, R.; Clardy, J. Atomic structure of the trypsin-A90720A complex: a unified approach to structure and function. *Chem. Biol*. **1994**, *1*, 113−117.

(67) Sandler, B.; Murakami, M.; Clardy J. Atomic Structure of Trypsin-Aeruginosin 98B Complex. *J. Am. Chem. Soc*. **1998**, *120*, 595−596.

(68) References to the Brookhaven PDB and additional descriptors are available as Supporting Information.

(69) Figures illustrating the results of the CP-ANN model for the training−test set are available as Supporting Information.

(70) Stubbs, M. T.; Reyda, S.; Dullweber, F.; et al. pH-Dependent binding modes observed in trypsin crystals: lessons for structure-based drug design. *ChemBioChem* **2002**, *3*, 246−249.

(71) Davis, A. M.; Teague, S. J. Hydrophobic Interactions and Failure of the Rigid Receptor Hypothesis. *Angew. Chem*. **1999**, *38*, 736−749.

(72) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results *J. Med. Chem*. **2004**, *47*, 2743−2749.

CI0401017