JCTC Journal of Chemical Theory and Computation

Semiempirical Comparative Binding Energy Analysis (SE-COMBINE) of a Series of Trypsin Inhibitors

Martin B. Peters and Kenneth M. Merz, Jr.*

Department of Chemistry, 104 Chemistry Building, The Pennsylvania State University, University Park, Pennsylvania 16802

Received November 21, 2005

Abstract: A scheme to decompose the intermolecular interaction energy of a series of complexes at the semiempirical (SE) level has been developed and validated. The comparative binding energy analysis (COMBINE) (Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. *J. Med. Chem*. **1995**, *38*, 2681−2691) and the semiempirical quantum mechanical method pairwise energy decomposition (PWD) (Raha, K.; van der Vaart, A. J.; Riley, K. E.; Peters, M. B.; Westerhoff, L. M. Kim, H.; Merz, K. M., Jr. *J. Am. Chem. Soc*. **2005**, *127*, 6583−6594) were coupled together to form SE-COMBINE. This approach calculates the residue pairwise electrostatic interaction energies, and QSAR models were built with the energies as descriptors using partial least squares (PLS). The application of SE-COMBINE was used as an investigation of the intermolecular interactions between 88 benzamidine inhibitors and trypsin and to test the ability of this new method to predict binding free energies. The predictive capability of SE-COMBINE is shown to be comparable to those of other QSAR methods, and using graphical intermolecular interaction maps (IMMs) enhances the interpretability of receptor-based QSARs.

Introduction

Prediction of the binding free energy of a ligand to a receptor is an unsolved problem. The answer to this problem is to develop a fundamental understanding of receptor−ligand interactions. The accurate prediction of binding free energies requires an exact energy function and a reliable conformational search method that can find the correct binding mode.[1] Considerable research has been carried out in these areas; however, the optimum compromise between computational efficiency and accuracy has yet to be reached.

Computational medicinal chemistry has taken a two-prong approach in the development of new drugs. First, virtual screening procedures, such as the computer-aided structure-based design (CASD) and simple counting methods, are used to screen virtual libraries of 10^6−10^9 molecules. CASD uses docking and scoring to predict the binding mode and affinity of new compounds. Docking methods have been shown to reproduce the binding modes within 2 Å of the crystal structure of protein−ligand complexes.[1] The CASD approach relies on the speed of the scoring function to rapidly evaluate each pose that is generated by docking. The second approach taken by computational chemistry is lead optimization. These methods are routinely carried out using Quantitative Structure Activity Relationship[2] (QSAR) approaches. Most QSAR methods are not receptor-based methods; in other words the receptor is not accounted for in model building. Indeed, this may be the only option if the receptor structure is unknown.

The widely used Comparative Molecular Field Analysis[3] (CoMFA) approach is a grid-based method where molecular properties such as steric (Lennard-Jones) and electrostatic (Coulomb) interactions are calculated between a probe atom and each molecule in the data set at every grid point. The properties at each grid point become descriptors, and models are built using multivariate techniques.

Receptor-based QSAR methods include COMparative BINding Energy analysis[4,5] (COMBINE) and MM/PBSA.[6,7] COMBINE uses a Molecular Mechanics (MM) potential energy function to calculate the intermolecular interactions between the receptor and ligand and builds QSAR models using multivariate statistical tools such as partial least squares (PLS).[8,9]

* Corresponding author e-mail: merz@qtp.ufl.edu. Present address: Department of Chemistry, Quantum Theory Project, University of Florida, 2328 New Physics Building, P.O. Box 118435, Gainesville, FL 32611-8435.

10.1021/ct050284j CCC: $33.50 © 2006 American Chemical Society
Published on Web 02/14/2006

The most accurate intermolecular interactions can be obtained using quantum mechanics (QM) methods. High-level QM methods such as Hartree−Fock (HF) and Density Functional Theory (DFT) are frequently used to study small organic systems and protein active sites; however, their use to study protein−ligand interactions is limited due to the high computational cost. Semiempirical (SE) QM methods were developed in the 1970s to reduce the computational cost with minimum loss in accuracy.[10] The most popular SE methods used today are based on the neglect of diatomic differential overlap (NDDO) approximation. The NDDO approximation reduces the number of integral evaluations in QM and in doing so changes the bottleneck of such methods to matrix diagonalization. The divide-and-conquer (D&C),[11−14] density matrix minimization,[15] and localized molecular orbital[16] methods have been developed to address the problem of matrix diagonalization enabling the application of SE methods to macromolecular systems. The D&C approach has been implemented in the program DivCon[17] which uses the SE Hamiltonians AM1,[18] PM3,[19,20] MNDO/d,[21,22] and PM3-PDDG.[23] Recently, D&C methods such as QMSCORE,[24,25] pairwise energy decomposition (PWD),[26] and DCNMR[27] have been developed to study protein−ligand interactions in DivCon. QMSCORE is a SE based score function that outperforms other score functions such as AutoDock and DrugScore. The PWD method is a novel approach where the electrostatic interaction energy is partitioned into self- and cross components between atoms. PWD has successfully been used to investigate the effect of binding of a series of fluorine-substituted ligands to human carbonic anhydrase II. DCNMR has been shown to predict NMR chemical shifts from the 3D structure of protein−ligand complexes.

In this work the PWD method was coupled to the COMBINE method creating SE-COMBINE to study a large set of protein−ligand complexes at the SE level of theory. PWD calculates the pairwise electrostatic interactions between a protein and ligand using the linear-scaling D&C approach. Similar to the COMBINE method, the residue pairwise energies were used to build QSAR models. The utility of SE-COMBINE was demonstrated by investigating the structure−activity relationship of a series of trypsin-like serine protease inhibitors.

Serine proteases are involved in many processes in the body such as protein digestion and blood coagulation.[28] The serine protease family of enzymes catalyzes protein hydrolysis. Trypsin, chymotrypsin, and elastase are common enzymes involved in the digestion of dietary proteins. Thrombin, factor Xa, and plasmin are key enzymes of the blood-clotting cascade. They differ only by their selectivity; for example, trypsin regiospecifically hydrolyzes at the carboxyl side of lysine and arginine amino acids, whereas chymotrypsin cleaves at aromatic sites. All serine proteases contain the catalytic triad Asp102, His57, and Ser195, which allows the catalytic cleavage of peptide bonds through an acyl intermediate as shown in Figure 1. The neighboring aspartic acid and histidine residues modify the serine from a hydroxyl to an alkoxide allowing the nucleophilic attack of the carbonyl group to occur.[29] The development of inhibitors of trypsin-like serine proteases has been an active area of research because they are important targets in the blood-clotting cascade and also serve as a useful model system to study protein−ligand interaction.

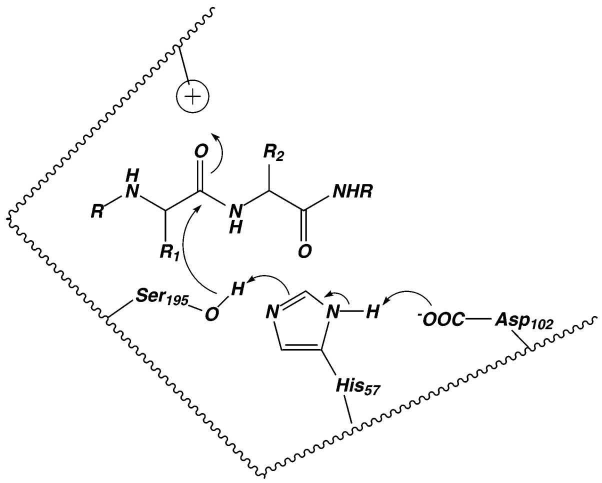


Figure 1. Trypsin hydrolysis mechanism. The general acid/base catalysis takes place using the catalytic triad of Asp102, His57, and Ser195. Asp102 removes a proton from His57, which activates Ser195 from a hydroxyl to an alkoxide nucleophile [adapted from Silerman, 2002].

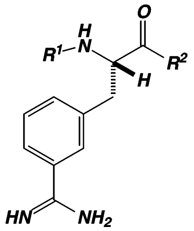


Figure 2. The structure of the 3-amidinophenylalanine molecule. Structural changes occur at two positions, R_1 and R_2 [adapted from Böhm et al.].

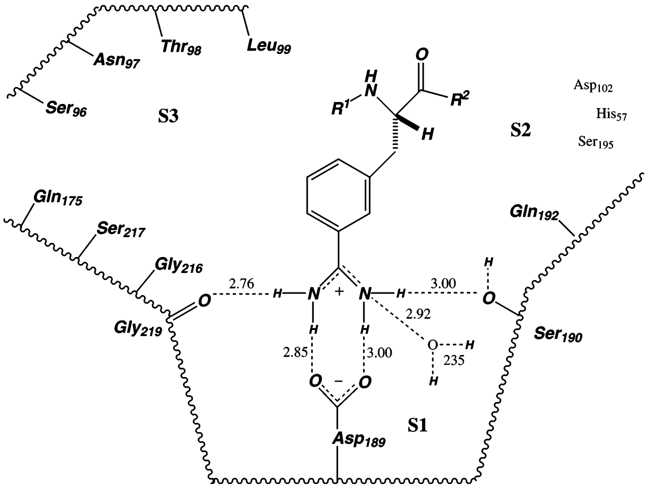


Figure 3. Schematic representation of 3-amidinophenylalanine bound to trypsin. The distances shown are determined from the complex of 3-TAPAP (Brookhaven Protein Data Bank reference: 1PPH) where R^1 is tosyl and R^2 is piperidine. Distances shown are in angstroms [adapted from Böhm et al.].

Table 1: 88 Trypsin Inhibitors[a]

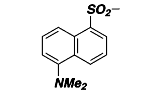
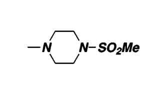
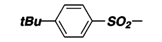
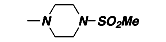
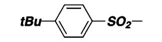
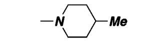
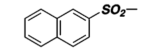
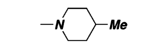
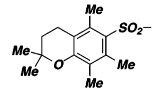
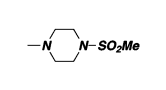
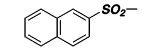
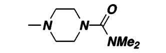
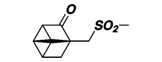
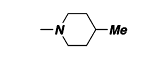
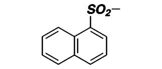
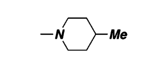
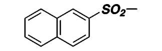
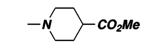
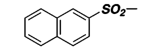
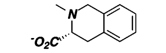
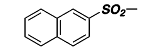
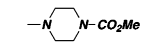
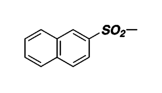
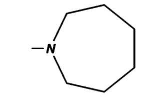
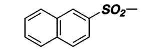
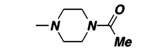
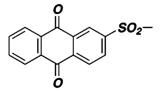
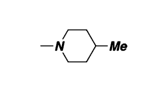
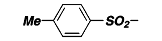
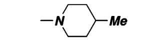
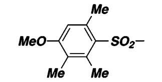
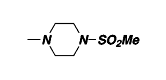
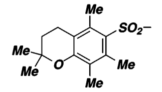
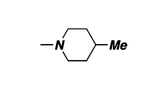
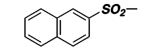
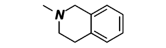
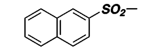
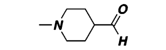
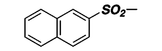
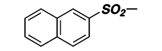
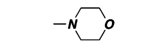
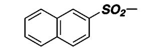
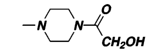
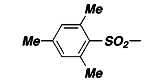
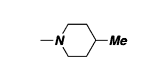
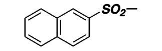
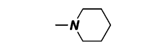
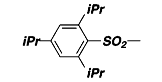
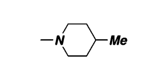
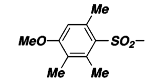
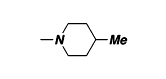
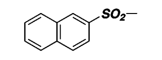
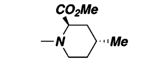
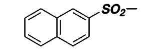
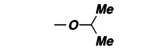
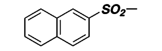
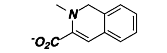
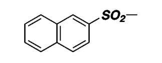
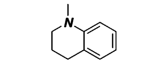
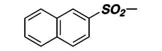
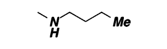
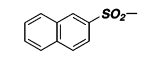
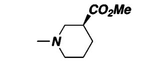
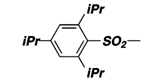
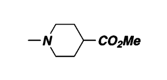
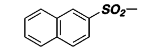
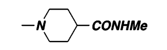
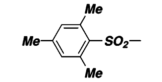
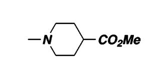
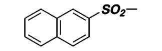
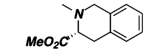
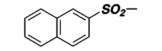
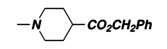
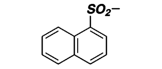
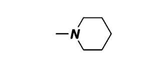
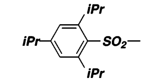
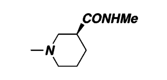
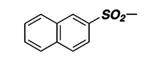
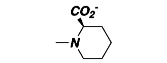
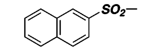
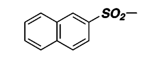
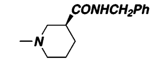
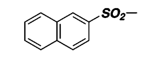
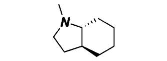
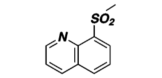
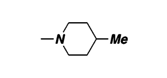
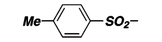
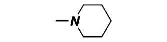
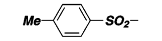
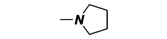
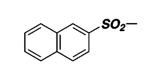
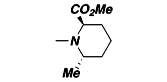
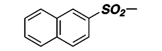
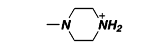
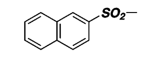
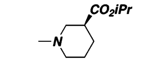
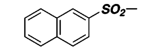
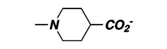
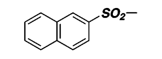
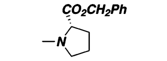
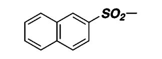
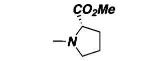
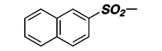
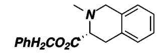
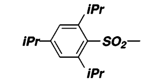
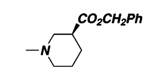
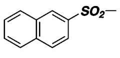
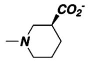
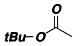
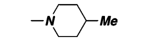
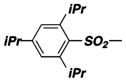
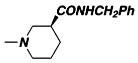
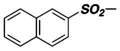
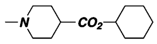
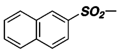
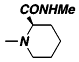
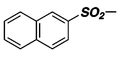
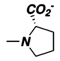
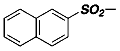
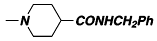
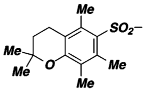
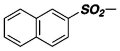
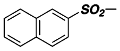
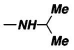
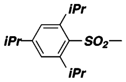
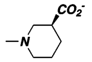
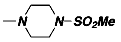
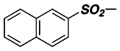
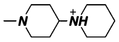
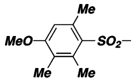
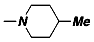
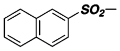
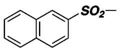
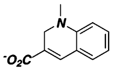
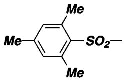
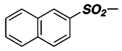
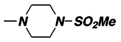
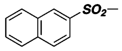
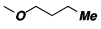
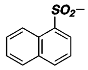
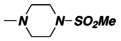
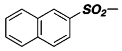
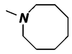
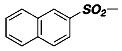
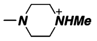
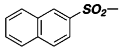
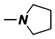
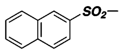
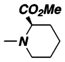
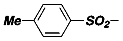
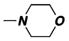
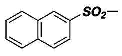
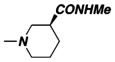
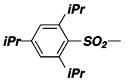
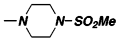
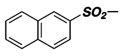
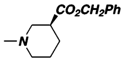
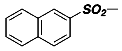
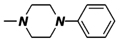
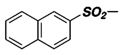
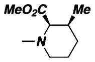
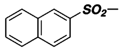
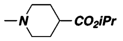
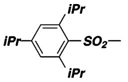
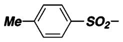
No.	R^1	R^2	Charge	pK$_i$
1			+1	6.770
2			+1	6.796
3			+1	6.699
4			+1	6.854
5			+1	6.119
6			+1	6.770
7			+1	6.201
8			+1	6.201
9			+1	7.444
10			0	6.886
11			+1	7.699
12			+1	6.260
13			+1	6.854
14			+1	7.131
15			+1	6.284
16			+1	5.745
17			+1	6.137
18			+1	6.585
19			+1	6.658
20		—OCH$_2$Ph	+1	6.284
21			+1	6.678
22			+1	5.959
23			+1	5.398
24			+1	6.481
25			+1	6.161
26			+1	6.108
27			+1	5.658
28			+1	5.854
29			0	5.347
30			+1	5.824
31			+1	5.398
32			+1	6.409
33			+1	6.569
34			+1	6.796
35			+1	6.004
36			+1	5.921
37			+1	7.174
38			+1	6.215
39			+1	6.102
40			0	6.201
41		—OMe	+1	5.602
42			+1	6.921
43			+1	5.921
44			+1	5.444
45			+1	5.921
46			+1	5.658
47			+1	5.678
48			+2	6.658
49			+1	6.367
50			0	6.237
51			+1	6.000
52			+1	5.092
53			+1	5.921
54			+1	6.071

Table 1: (Continued)

No.	R^1	R^2	Charge	pK_i
55			0	6.357
56			+1	4.854
57			+1	6.337
58			+1	7.097
59			+1	5.102
60			0	4.796
61			+1	6.569
62		$-NHMe$	+1	4.495
63		$-NHMe$	+1	4.602
64			+1	4.796
65			0	5.620
66	H—		+2	4.538
67			+2	6.000
68		$-NHMe$	+1	3.854
69	H—		+2	4.538
70		$-CO_2^-$	0	3.928
71			0	4.509
72		$-NHMe$	+1	3.000
73			+1	6.721
74			+1	6.585
75			+1	6.495
76			+1	6.215
77			+2	5.886
78			+1	6.357
79			+1	5.721
80			+1	6.149
81			+1	6.509
82			+1	6.009
83			+1	6.796
84			+1	7.569
85			+1	5.745
86			+1	7.638
87		$-NHMe$	+1	4.585
88		$-NHMe$	+1	4.337

[a] Substituents at positions at R^1 and R^2, formal charges and pK_i values are listed [adapted from Böhm et al.]

Trypsin is synthesized in the pancreas as a zymogen (inactive enzyme) called trypsinogen. When required, trypsinogen is secreted into the small intestine through the bile duct and after enzymatic removal of an N-terminal amino acid sequence trypsin (24kDa) is formed. Trypsin has a large binding pocket, **S1**, adjacent to the catalytic site with an aspartic acid at the base. This pocket favors the binding of the positively charged amino acids, lysine and arginine. The strong ionic interaction allows for the cleavage reaction to take place. The enzymes thrombin and factor Xa have similar **S1** pockets; however, the **S2** and **S3** pockets vary in composition and in size (thrombin has the insertion loop, Tyr60A-Trp60D, whereas the others do not).[30] Thus the development of selective inhibitors of trypsin, thrombin, and factor Xa has posed a challenge for both experimental and computational research.[31]

Since 1965, benzamidine-based inhibitors of trypsin-like proteases have been developed.[32] The amidinophenylalanine group mimics the guanidinalkyl functional group of arginine as shown in Figure 2. The X-ray structure of $N\alpha$-[4-toluene sulfonyl]-L-*m*-amidino-phenylalanyl (3-TAPAP, a 1.2 μmol/L inhibitor) bound to trypsin (1.9 Å resolution) was reported in 1991 (PDB: 1PPH).[30] A schematic representation of the key interactions between the 3-amidinophenylalanine group and the **S1** pocket of trypsin from 1PPH is shown in Figure 3. The amidino group forms a near symmetric salt bridge with Asp189 and is also hydrogen bonded to both Gly219@O and a water molecule. The phenyl ring is sandwiched between the sequences Ser190-Gln192 and Trp215-Gly216. Gly216 forms hydrogen bonds with the amino and carbonyl group of the *m*-amidino-phenylalanine moiety. The tosyl group, R^1, fills the **S3** pocket and lies perpendicular to the indole group of Trp215, while anoxygen of the sulfonyl group points toward Gly219@N. The piperidine group occupies the **S2** pocket and is flanked on either side by His57 and the toluene group of the tosyl moiety. Using the benzamidine scaffold, a series of inhibitors was reported,[33,34] and recently 3D QSAR techniques such as CoMFA[35], CoMSIA,[36] and QSM[37] have been used to investigate the inhibitor selectivity between thrombin, trypsin, and factor Xa.[38]

Trypsin and its inhibitors are very well characterized, i.e. protein−ligand complex structures and binding affinity data are available, thus providing an excellent starting point for a computational study. The semiempirical quantum mechanical decomposed intermolecular interactions between trypsin and a series of inhibitors shown in Table 1 were examined in this study. Using the protein-residue—ligand-fragment interaction energies, a comparative binding energy analysis was carried out using PLS to build a receptor-based 3D-QSAR model.

Computational Approach

Consider the interaction of a receptor R, with a ligand L, to form the complex R ·L:

$$\mathrm{R} + \mathrm{L} \rightleftharpoons \mathrm{R{\cdot}L} \tag{1}$$

The interaction energy can be calculated using the following

$$E_{\mathrm{INT}} = E_{\mathrm{R\cdot L}} - (E_{\mathrm{R}} + E_{\mathrm{L}}) \tag{2}$$

where $E_{\mathrm{R\cdot L}}$ is the energy of the complex, and E_{R} and E_{L} are the energies of the receptor and ligand, respectively. Equation 2 can also be represented with ΔE_{R} and ΔE_{L} as the change in energy of the receptor and ligand upon binding as in eq 3.

$$E_{\mathrm{INT}} = E_{\mathrm{R\cdot L}} + \Delta E_{\mathrm{R}} + \Delta E_{\mathrm{L}} \tag{3}$$

Carrying out a residue-based pairwise decomposition of the interaction energy leads to the following

$$E_{\mathrm{INT}} = \sum_I \sum_J E_{IJ} + \sum_I \sum_{K<I} \Delta E_{IK} + \sum_J \sum_{L<J} \Delta E_{JL} + \sum_I \Delta E_I + \sum_J \Delta E_J \tag{4}$$

where I is the index of the residues in the receptor, J is the index of fragments in the ligand, E_{IJ} is the residue(receptor)-fragment(ligand) interaction energy (a true cross term, this term is only present in the complex), ΔE_{IK} is the change in the interresidue energy upon binding of the residues in the receptor, ΔE_{JL} is the change in the interfragment energy upon binding of the fragments in the ligand, ΔE_{I} is the change in intraresidue energy upon binding of the residues in the receptor, and ΔE_{J} is the change in intrafragment energy upon binding of the fragments in the ligand.

Considering the above in terms of a classical molecular mechanics force field, the first term would be the electrostatic and van der Waals interactions between receptor residues and ligand fragments. The second and third terms would be change in electrostatic and van der Waals interactions between receptor residues and other residues and ligand fragments and other fragments upon complexation. The fourth and fifth terms would be the change in the bond, angle, torsion, and nonbonded interactions of receptor residues and ligand fragments.

Within a semiempirical approach the binding energy expression of eq 4 can be expressed in terms of the quantities derived by Raha et al.[26]

$$\begin{aligned}
E_{\mathrm{INT}} = {} & \sum_I \sum_J \Big(\sum_A \sum_B E_{AB} + E'_{AB} + E^{core}_{AB}\Big) + A \in I, B \in J \\
& \sum_I \sum_{K<I} \Big(\sum_A \sum_B \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\Big) + A \in I, B \in K \\
& \sum_J \sum_{L<J} \Big(\sum_A \sum_B \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\Big) + A \in J, B \in L \\
& \sum_I \Big(\sum_A \{\Delta E_A + \sum_{B<A} \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\}\Big) + \mathrm{A,B} \in \mathrm{I} \\
& \sum_J \Big(\sum_A \{\Delta E_A + \sum_{B<A} \Delta E_{AB} + \Delta E'_{AB} + \Delta E^{core}_{AB}\}\Big)\ A,B \in \mathrm{J}
\end{aligned} \tag{5}$$

where

$$E_A = \frac{1}{2} \sum_{\mu}^{A} \sum_{\nu}^{A} P^{AA}_{\mu\nu} \left(2H^{AA}_{\mu\nu} + \sum_{\lambda\sigma}^{A} P^{AA}_{\lambda\sigma} \left[(\mu^A \nu^A | \sigma^A \lambda^A) - \frac{1}{2} (\mu^A \sigma^A | \lambda^A \nu^A) \right] \right) \tag{6}$$

$$E'_{AB} = \sum_{\mu}^{A} \sum_{\nu}^{A} \sum_{\lambda\sigma}^{B} P^{AA}_{\mu\nu}\, P^{BB}_{\lambda\sigma} (\mu^A \nu^A | \sigma^B \lambda^B) \tag{7}$$

$$E_{AB} = \sum_{\mu}^{A} \sum_{\nu}^{B} P^{AB}_{\mu\nu} \left(2H^{AB}_{\mu\nu} - \frac{1}{2} \sum_{\lambda}^{B} \sum_{\sigma}^{A} P^{BA}_{\lambda\sigma} (\mu^A \sigma^A | \lambda^B \nu^B) \right) \tag{8}$$

$$E^{core}_{AB} = \sum_A \sum_{B<A} \frac{Z_A Z_B}{R_{AB}} \tag{9}$$

PWD calculates the self-energy of the atom, E_A, core—electron interactions, E^{core}_{AB}, electron—electron repulsions, E'_{AB}, and exchange between the atoms, E_{AB}, as shown in eqs 6—9. H is the one-electron matrix, F is the Fock matrix, and P is the density matrix. Z is the nuclear charge on an atom, R_{AB} is the atomic separation between A and B. Equation 5 represents the decomposition of the semiempirical interaction energy between a receptor and ligand. The E_A term has a large negative energy contribution to the total energy since it contains the one-center terms. E'_{AB} contains all the electronic repulsion, and so it is a positive contributor to the energy which comes from the diagonal block of the Fock matrix. E_{AB} contains the exchange repulsion between atoms and is a small negative contributor to the total energy, which stems from the off-diagonal elements of the Fock, one-electron, and density matrices. As originally described, it contains most of the binding energy. Accepting, as an approximation, that the receptor and ligand conformations remain the same upon binding, the decomposed energy becomes

$$\begin{aligned}
E_{\mathrm{INT}} = {} & \sum_I \sum_J (E_{AB} + E'_{AB} + E^{core}_{AB}) + \\
& \sum_I \sum_{K<I} (\Delta E_{AB} + \Delta E'_{AB}) + \sum_J \sum_{L<J} (\Delta E_{AB} + \Delta E'_{AB}) + \\
& \sum_I (\Delta E_A + \Delta E_{AB} + \Delta E'_{AB}) + \sum_J (\Delta E_A + \Delta E_{AB} + \Delta E'_{AB})
\end{aligned} \tag{10}$$

Equation 10 as written implies a sum over A and B in the first 3 terms, over A in the fourth term, and over B in the final term.

Mol	Act	IJ-E'_{AB}	IJ-E_{AB}	IJ-E^{core}_{AB}	IK-E'_{AB}	IK-E_{AB}	JL-E'_{AB}	JL-E_{AB}	I-E'_{AB}	I-E_{AB}	I-E^{core}_{AB}	J-E'_{AB}	J-E_{AB}	J-E^{core}_{AB}
1	•	·	·	·	·	·	·	·	·	·	·	·	·	·
2	•	·	·	·	·	·	·	·	·	·	·	·	·	·
g	•	·	·	·	·	·	·	·	·	·	·	·	·	·
N	•	·	·	·	·	·	·	·	·	·	·	·	·	·

Figure 4. Schematic diagram of an example data table used in SE-COMBINE. The size of the descriptor matrix is defined by the number of compounds, *N*, and by the number of descriptors. The experimental data, Act, is a single column in the data table, while the indices I, J, K, and L refer to eq 10. IJ − E'_{AB}, IJ − E_{AB}, and IJ − E^{core}_{AB} are energy terms between receptor residues and ligand fragments, true cross terms. IK − E'_{AB} and IK − E_{AB}, are energy terms between pairs of receptor residues. JL − E'_{AB} and JL − E_{AB} are energy terms between pairs of ligand fragments. The remaining terms are residue and fragment self-energy terms.

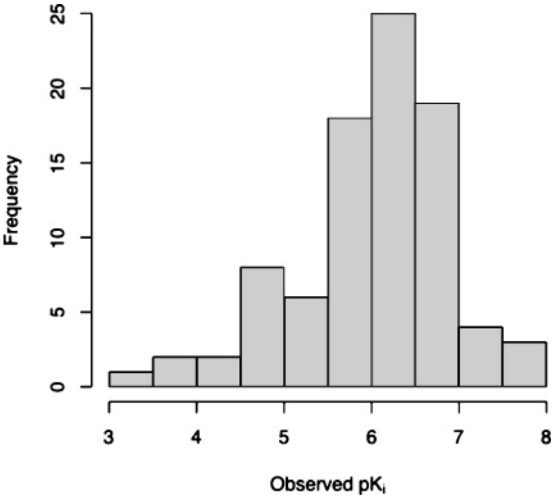


Figure 5. Trypsin inhibitor activity frequency distribution. Affinities spread over a 4.7 logarithm unit's range, which allows a statistically significant 3D QSAR to be derived.

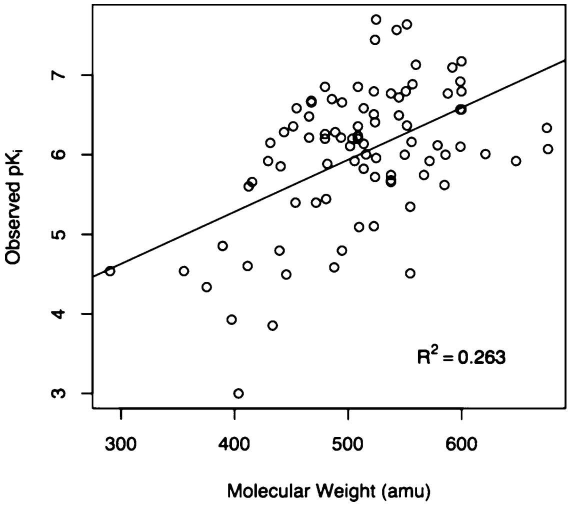


Figure 6. Trypsin inhibitors pK_i versus molecular weight. Poor correlation is observed with an R^2 value of 0.26.

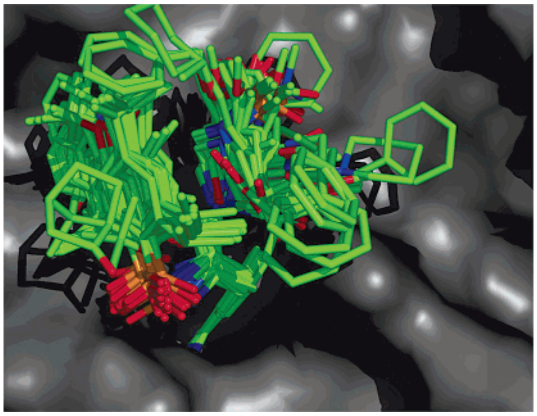


Figure 7. Inhibitors aligned in the active site of trypsin. Trypsin is represented as a surface and inhibitors as sticks. Hydrogen atoms are not shown for clarity [adapted from Böhm et al.].

The pwdPy program was developed in order to perform a pairwise decomposition of the interaction energy between the ligand fragments and the protein residues using the formalism described above in eq 10. That is, the program was used to read the DivCon output of the ligand, protein, and complex calculations and to produce a descriptor table similar to the one shown in Figure 4.

Procedure

(1) Data Set. The crystal structure of 3-TAPAP bound to trypsin (1PPH) was obtained from the Protein Data Bank (PDB). 3-TAPAP is a 3-amidinophenylalanine based inhibitor of trypsin. Eighty-eight compounds (coordinates and activity data), including 3-TAPAP (all fully protonated), which bind to trypsin, were kindly provided by Prof. Gerhard Klebe. The L-conformations of the central phenylalanine are more potent by a factor of 50−100 over the D-conformations; however, the pK_i reported are mixtures of the L and D forms.[34] Only the L-conformations of the compounds were used in this study. All 88 structures share a common core and differ at positions R^1 and R^2 as shown in Figure 2. The structures of the R^1 and R^2 groups for all compounds and their activity data are listed in Table 1. The affinities of the inhibitors spread over a range of 4.7 pK_i units. However, as shown in Figure 5 the majority of the inhibitors' affinities lie in the range between 5.5 and 7.0 pK_i units. Note that the variation in size of the R groups does not translate to a molecular weight dependence on binding affinity as shown in Figure 6.

(2) Molecular Mechanics Modeling of the Receptor. The 3-TAPAP structure and all water molecules except number 235 were removed from the 1PPH crystal structure. All

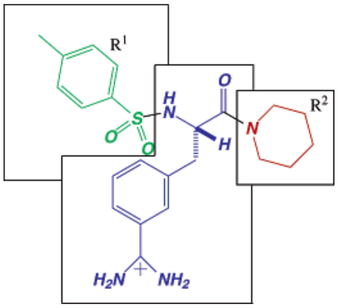


Figure 8. Schematic diagram of a trypsin inhibitor fragmentation. The structure in blue is the 3-amidino-phenylalanine moiety (**APM**). The **TOS** group is colored green, while the **PIP** group is shown in red.

references to the amino acid names and numbers follow that used in 1PPH. Wat235 is the characteristic water molecule present in the **S1** pocket. Hydrogen atoms were added to the protein using the LEaP module of AMBER, followed by a hydrogen minimization (1500 steps) using the SANDER module of AMBER 8.[39] All acidic residues were assumed to be deprotonated while all basic residues were protonated.

(3) Molecular Mechanics Modeling of the Complexes. The 88 ligands were aligned onto 3-TAPAP (**45**) using the 3-amidinophenylalanine moiety as a template. Each ligand was placed in the active site of 1PPH as shown in Figure 7. The 88 compounds vary in size and shape, and so some close contacts were expected. To correct this the inhibitors were allowed to relax in the active site using a restrained minimization of 1500 steps (500 steepest descent followed by 1000 conjugate gradient), followed by a full minimization of all atoms in the system (500 steepest descent followed by 1000 conjugate gradient steps) using AMBER.

(4) Semiempirical D&C Calculations. Semiempirical D&C calculations were performed using the PM3 Hamiltonian within the program DivCon. Five calculations were performed for each of the 88 compounds in the data set: (1) protein only, (2) ligand (1 fragment), (3) ligand (3 fragments), (4) complex (ligand with 1 fragment), and (5) complex (ligand with 3 fragments). The protein was divided into subsystems based on the standard amino acid residue definitions. All atoms of the ligand were grouped into one fragment in calculation 2, and the fragment name was set to **TAP**. Each ligand in the data set was also divided into three groups or fragments, e.g. **45** is shown in Figure 8. The first fragment consists of the 3-amidinophenylalanine moiety (**APM**), the R^1 group contained aryl sulfonyl groups (**TOS**), and the third contains either piperidine or piperazine groups (**PIP**). The fragments were named based on those residues found in 3-TAPAP of 1PPH. A cutoff for the Fock matrix of 20 Å and a divide-and-conquer buffering scheme of 4.2/2 Å were used throughout the entire study. Note that the total

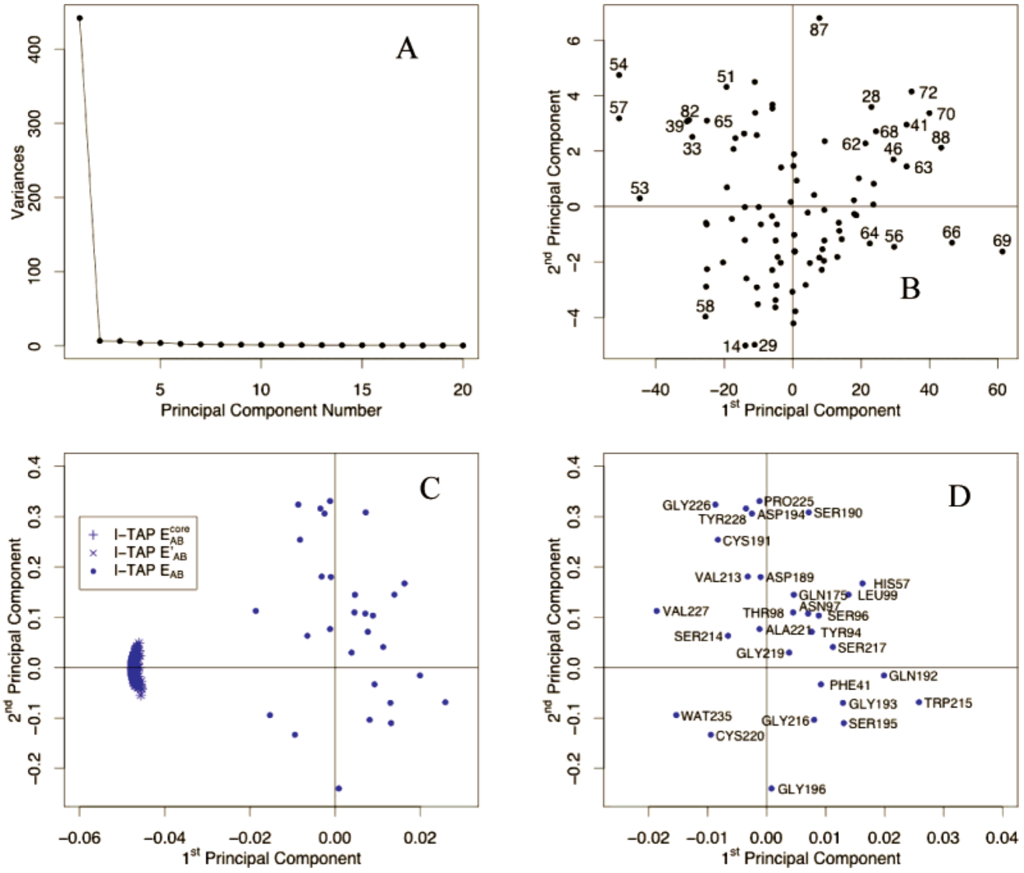


Figure 9. Model Lig1C PCA results. (A) Scree plot. (B) Score plot of PC 1 versus 2. Points representing complexes of interest are labeled using ligand numbers. (C) Loading plot of PC 1 versus 2. The types of descriptors in this model are shown in the legend where I denotes any amino acid in the protein. (D) Loading plot of PC 1 versus 2 where only E_{AB} descriptors are considered. Labels shown are the protein residue name and number involved in the interaction with the ligand.

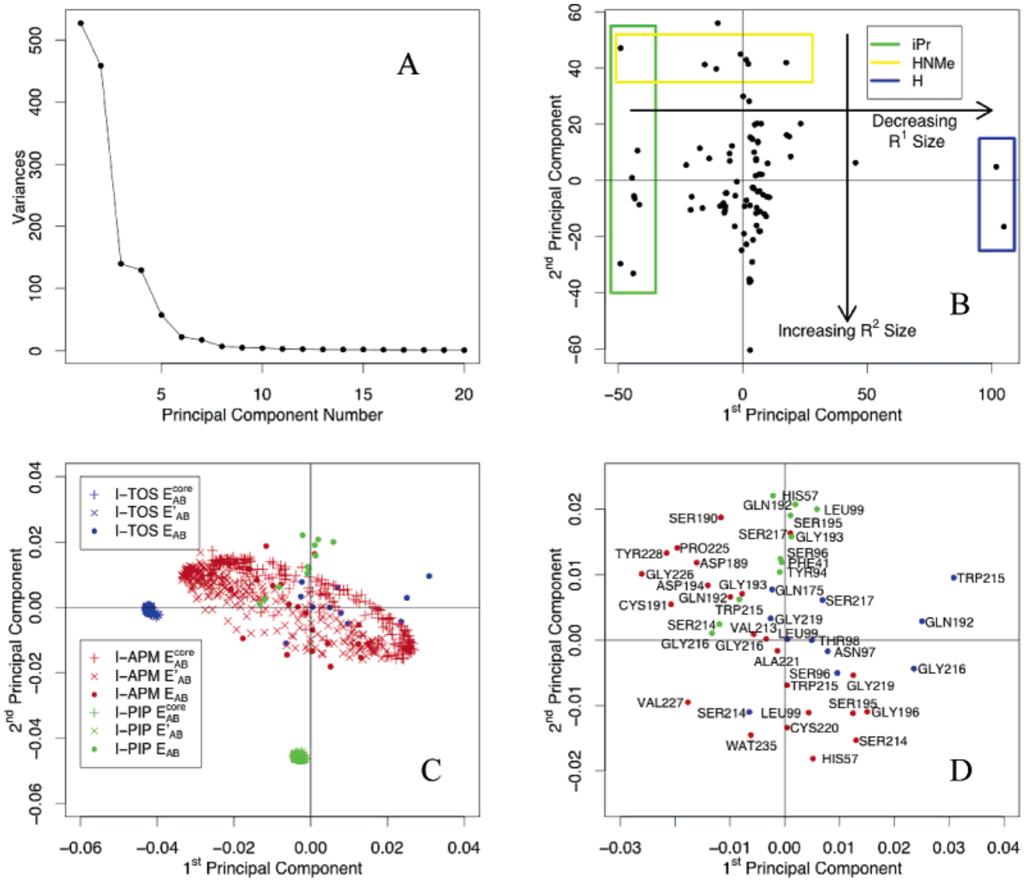


Figure 10. Model Lig3C PCA results. (A) Scree plot. (B) Score plot of PC 1 versus 2. Points representing complexes of interest are labeled using ligand numbers. Highlighted are isopropyl, ^{i}Pr, and hydrogen, H, on R^1 and HNMe on R^2. (C) Loading plot of PC 1 versus 2. The types of descriptors in this model are shown in the legend where I denotes any amino acid in the protein. (D) Loading plot of PC 1 versus 2 where only E_{AB} descriptors are considered. Descriptors are labeled by protein residue, whereas the colors represent the fragments of the ligand.

interaction energy between the receptor and ligand does not change after fragmentation.

(5) Chemometric Analysis. The pwdPy program was used to pairwise-decompose the interaction energy between the ligand fragments and the protein residues. The statistical analysis of the decomposed energies was preformed using the program R.[40]

As a first step, Principal Component Analysis (PCA) was carried out to examine the distribution of the complexes in the descriptor space. The similarity/dissimilarity between inhibitors was investigated using score plots. The descriptor pool was pruned to remove descriptors that returned zero values. Auto-scaling was applied to the descriptor matrices, or, in other words, each descriptor was processed to have a mean of zero and a standard deviation of one. This ensured that certain variables did not dominate due to their magnitude. PLS models were built to explore the structure−activity relationship of the inhibitors. Internal validation was carried out using leave-one-out (LOO) cross-validation, and the optimal dimensionality of each model was assigned from its cross-validated predictive ability. External validations were also carried out where 10 structures were removed randomly from the original data matrix to become the prediction set, while the rest remained as the training set. Ten such prediction sets were generated. Descriptor pruning and auto-scaling was applied to the training set following the above procedure. After the models were generated using the training set, the prediction set was autoscaled using the means and standard deviations from the training set. Another external validation study was carried out where the training and prediction sets were predefined by Böhm et al. This was used in order to compare SE-COMBINE to methods such as CoMFA, CoMSIA, and quantum similarity.

Together with the square of the Pearson's correlation coefficient, R^2, and the cross-validated correlation coefficient, Q^2, the standard deviation of error of calculations, SDEC, and the standard deviation of error prediction, SDEP, were used to assess the quality of the models. SDEP can also be defined as the root-mean-squared error of the dependent variables in a LOO scheme or external data set. Similarly, SDEC is calculated for those variables used to build the model or training set. For each model, the biological activities of the inhibitors were scrambled randomly, and the activities were predicted, as a way of detecting the possibility of chance correlation. And in all cases only negative Q^2 values were observed.

Results

The pairwise interactions between the 224 amino acid residues of trypsin and a water molecule with each inhibitor

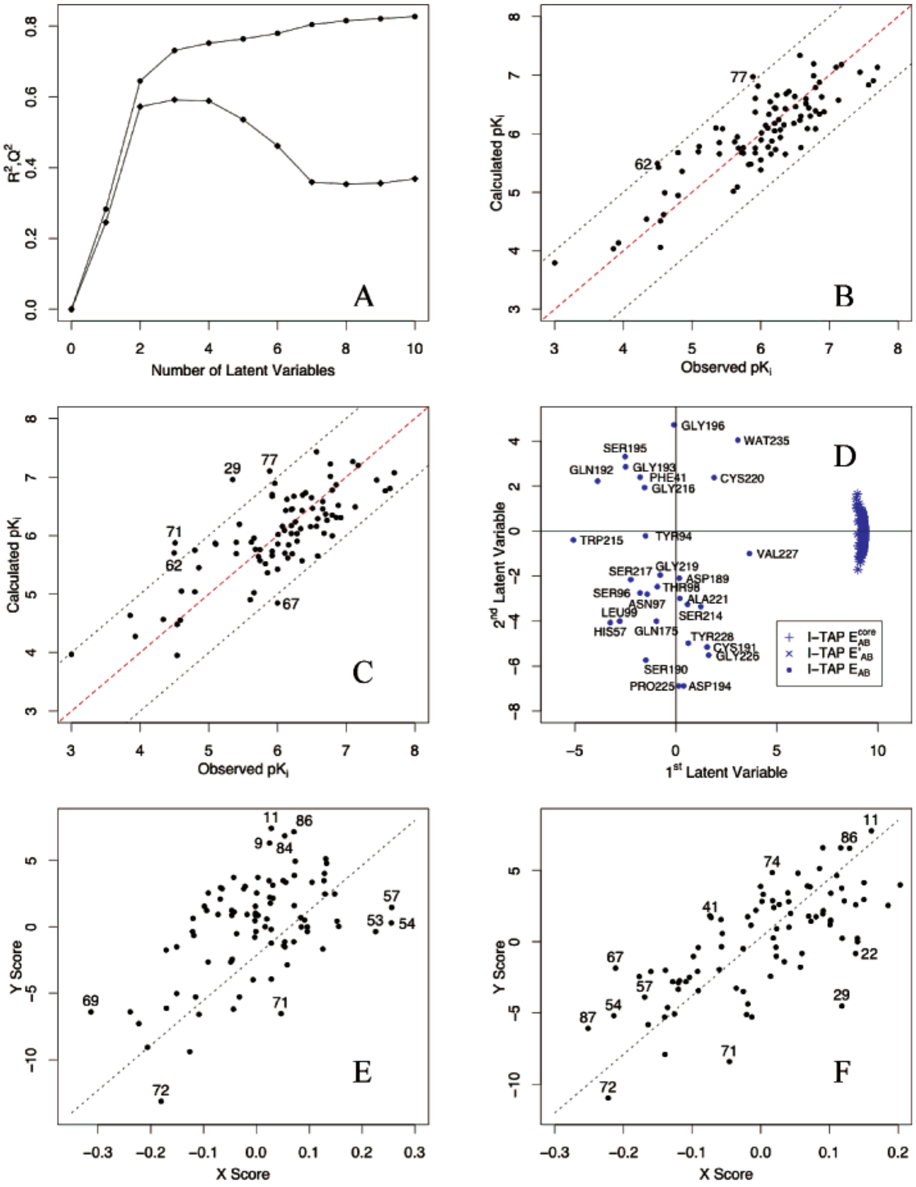


Figure 11. Model Lig1C PLS results. (A) R^2 and Q^2 versus number of latent variables. R^2 is represented as solid circles and Q^2 as solid diamonds. (B) Observed versus calculated pK_i values from internal validation. The red line represents the optimal correlation, while the blue lines are a pK_i unit from the optimal line. (C) Observed versus calculated pK_i values from LOO cross-validation. (D) Loading plot of the first and second latent variables. (E) X-score versus Y-score for the first latent variable. (F) X-score versus Y-score for the second latent variable.

were calculated. The effect of fragmentation of the ligand structure was investigated by considering a single (Lig1) and triple fragment (Lig3) scheme. The Lig1 scheme yields a total of 25 878 (model Lig1A) descriptors (computed using eq 11) to fully decompose E_{INT} (eq 10). Only considering the cross term E_{IJ}, this number reduces to 672 (model Lig1B). It was found that the majority of the E_{AB} terms were zero [Tests with a water dimer showed that E_{AB} is zero with oxygen−oxygen distances greater than 4 Å.]; therefore, a E_{AB} descriptor was removed if more than 95% of its terms were zero. The remaining descriptors were auto-scaled because the E_{AB}, E'_{AB}, and E^{core}_{AB} terms span different ranges (model Lig1C). Using the same procedure the Lig3 scheme produces 52 655 descriptors (model Lig3A). This number was reduced to 2016 when only E_{IJ} interactions were considered (model Lig3B). The E_{AB} terms were pruned reducing the dimensionality, and autoscaling was applied, thus producing model Lig3C. The total number of the descriptors per model is given in Table 2.

Table 2: Number of Descriptors per QM-COMBINE Model

	number of descriptors	
model	Lig1	Lig3
A	25878	52655
B	672	2016
C	477	1389

$$((I^*J) + I + J)^*3 + \sum_{n}^{I}(n-1)^*2 + \sum_{n}^{J}(n-1)^*2 \qquad (11)$$

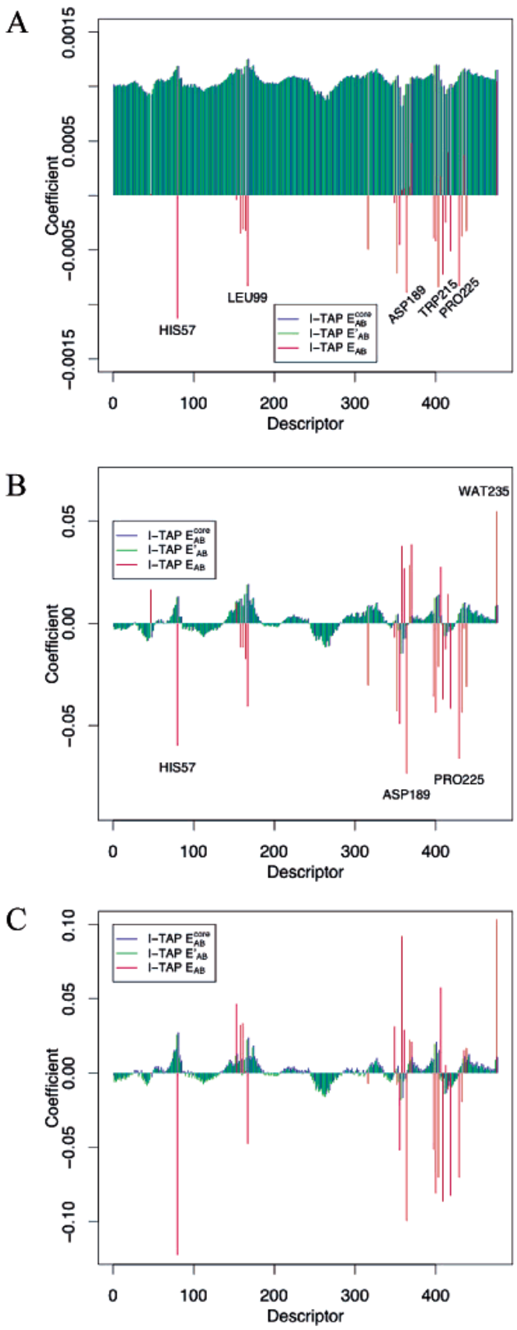


Figure 12. Model Lig1C PLS coefficient plots. (A) Latent variable 1. (B) Latent variable 2. (C) Latent variable 3.

Table 3: Models Lig1C and Lig3C PLS Results[a]

model	LV	*X* variance (cumulative)	R^2	Q^2	SDEC	SDEP	SDEC^{ext}	SDEP^{ext}
Lig1C	1	92.88	0.28	0.25	0.74	0.76	0.75	0.68
	2	94.18	0.65	0.57	0.57	0.57	0.46	0.47
	3	94.68	0.73	0.59	0.45	0.56	0.32	0.42
Lig3C	1	32.53	0.40	0.32	0.68	0.72	0.69	0.60
	2	45.21	0.50	0.42	0.62	0.67	0.62	0.58
	3	79.64	0.51	0.43	0.61	0.66	0.61	0.58
	4	82.86	0.63	0.48	0.53	0.64	0.51	0.59
	5	89.63	0.67	0.53	0.50	0.60	0.46	0.58
	6	94.76	0.70	0.55	0.48	0.59	0.42	0.55
	7	96.38	0.74	0.58	0.45	0.57	0.38	0.51

[a] LV represents the number of latent variables in the model. The optimal number of LVs is shown in bold. R^2 and Q^2 represent the correlation coefficient of training and LOO. SDEC and SDEP are the standard deviations of calculation and prediction for internal validation. SDEC^{ext} and SDEP^{ext} are similarly defined applied to the external validation (values reported are averages of the 10 prediction sets).

Table 4: Predefined Training and Prediction Set PLS Results[a]

model	LV	*X* variance (cumulative)	R^2	Q^2	SDEC	SDEP	R^{2ext}	SDEP^{ext}
Lig1C	1	94.84	0.29	0.25	0.74	0.76	0.27	0.77
	2	94.08	0.65	0.56	0.51	0.58	0.73	0.54
	3	94.74	0.75	0.58	0.44	0.57	0.64	0.57
Lig3C	1	33.54	0.37	0.27	0.69	0.75	0.50	0.67
	2	55.13	0.46	0.37	0.64	0.70	0.67	0.57
	3	80.05	0.50	0.39	0.62	0.69	0.64	0.59
	4	83.07	0.62	0.45	0.54	0.66	0.62	0.60
	5	87.57	0.68	0.49	0.49	0.62	0.66	0.58
	6	94.53	0.71	0.54	0.47	0.60	0.61	0.59
	7	96.29	0.75	0.55	0.43	0.59	0.63	0.56

[a] LV represents the number of latent variables in the model. The optimal number of LVs is shown in bold. R^2 and Q^2 represent the correlation coefficient of training and LOO. SDEC and SDEP are the standard deviations of calculation and prediction for internal validation. R^{2ext} and SDEP^{ext} are similarly defined applied to the external validation.

Table 5: Comparison between Various 3D-QSAR Methods and Models Generated by SE-COMBINE[a]

method	descriptors	Q^2	LV	compds	predictive R^2	SDEP
CoMFA	2184	0.63	5	72	0.65	0.52
CoMSIA	2184	0.75	9	72	0.84	0.35
MQS matrices	20/72	0.63	8	72	0.75	0.47
fragment QS-SM	15−25/95	0.69	8	69	0.92	0.51
SE-COMBINE Lig1C	477	0.58	3	72	0.64	0.57
SE-COMBINE Lig3C	1389	0.55	7	72	0.63	0.56

[a] The total number of descriptors in each model is given. The LOO cross-validated Q^2, number of latent variables (LV), compounds (compds) in the training set, predictive R^2 for the 16 compound prediction set, and the SDEP.

Model Lig3A contained 52 655 descriptors. This model could not be handled using current computer hardware (required over 4 GB of RAM) and was skipped. The same holds for model Lig1. Models Lig1B and Lig3B were statistically compromised since the majority of the E_{AB} terms were zero. Therefore the first model to be considered was Lig1C and then Lig3C.

Principal Component Analysis: Model Lig1C. A PCA of the descriptor matrix for the 88 complexes was performed using the autoscaled variables.[41,42] The scree plot of the PCA shown in Figure 9.A illustrates that two principal components (PCs) successfully models the data. PC 1 and 2 explain 94.2% of the variance in the *X* matrix. The score plot of the first and second PC is shown in Figure 9.B. The score plot is divided into quadrants; the upper right quadrant contains complexes that have positive scores for both PC 1 and PC 2. These complexes all have small R^2 groups, while the lower left quadrant has large R^2 groups. A similar trend is observed for the R^1 groups: the top left quadrant contains large R^1 groups, while the lower right contains complexes with smaller R^1 groups. The complexes are distinguished in the descriptor space, with R group size playing a relevant role.

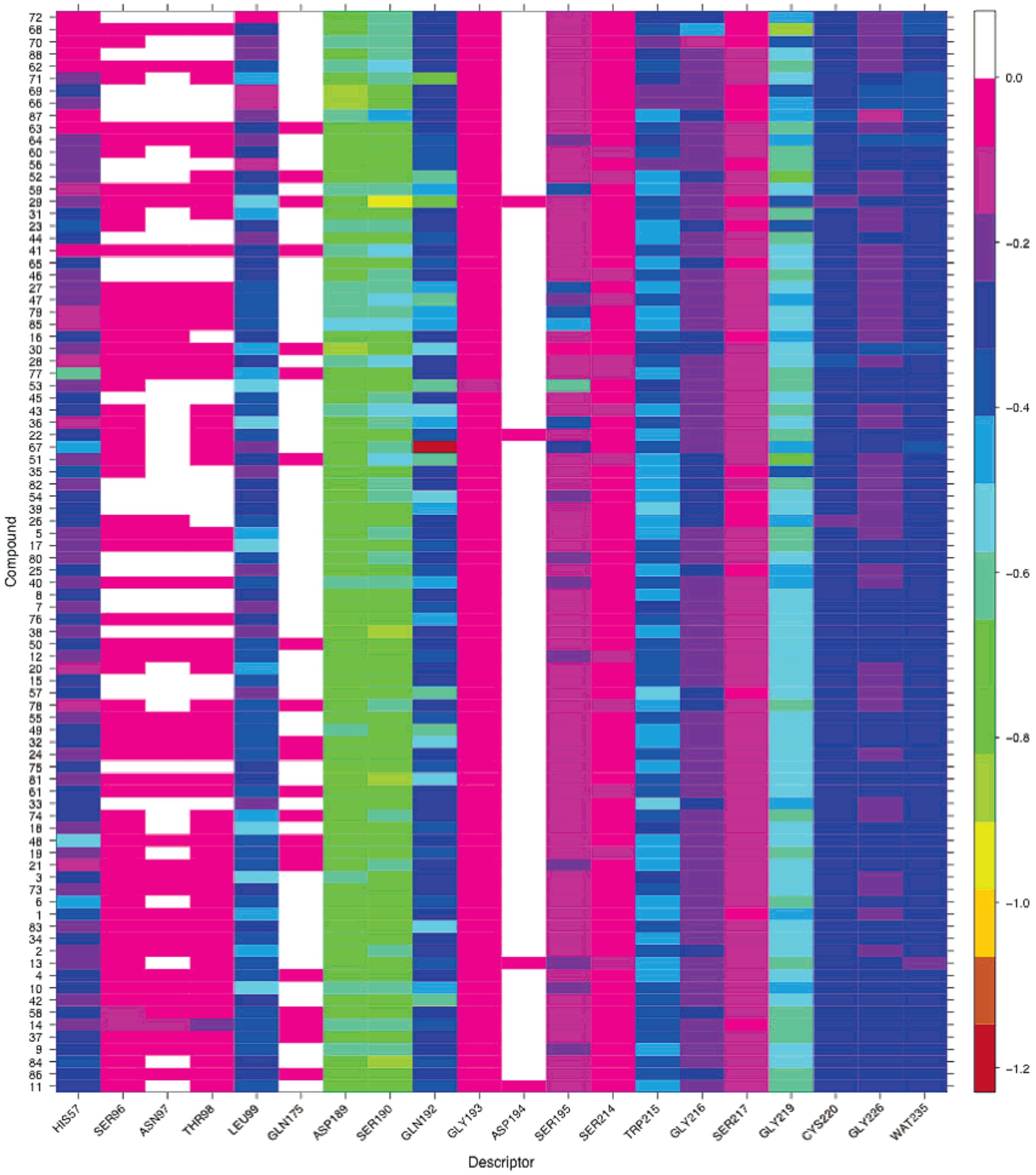


Figure 13. Model Lig1C intermolecular interaction map (IMM) of the important E_{AB} descriptors. The key residues of trypsin that interact with the single fragment ligand (**TAP**) label the *x*-axis. The compounds on the *y*-axis are ordered with respect to activity. The activity decreases from top to bottom. The legend indicates the magnitude of the unscaled descriptor in eV.

The loadings plot of PC 1 and PC 2 is shown in Figure 9.C. There are two clear clusters in the loading plot, the first contains the E_{AB}^{core} and E'_{AB} descriptors and the second all the E_{AB} terms. Importantly, the derivation of the PWD ascribes to E_{AB} the binding information between the fragments. Figure 9.D takes a closer look at the 29 E_{AB} terms. The scores and loadings in a PCA are related: the scores provide the coordinates of the data in the so-called hyper-planes, and the loadings present the direction of each dimension. The loading plot sheds light on the reason clustering in the score plot occurs, because the link between the two plots can be made by comparing the position of the original variables in the loading plot and the position of the compounds in the score plot. His57 is a part of the catalytic triad of trypsin and is prominent in the upper right quadrant of the loading plot. This is in agreement with the score plot where groups with smaller R^2 group were found in that region. The interaction of His57 and the inhibitors would be expected to be greater with a larger R group. Trp215 in the lower right quadrant dominates which is a key residue in the **S3** pocket. Inhibitors with smaller R^1 groups such as **66** and **69** are shown in the lower right quadrant. The residues that lie between the upper quadrants make up the base of the **S1** pocket, which interact with the benzamidine moiety of the inhibitors. There is no obvious relationship between their orientations in the loading plot and the score plot.

Principal Component Analysis: Model Lig3C. This model contains 1389 energy descriptors, and a PCA model was generated where 7 components account for 97% of the variation in the *X* matrix as shown in Figure 10.A. The score plot of PC 1 and 2 is shown in Figure 10.B with two components explaining 71% of the variance in the *X* matrix.

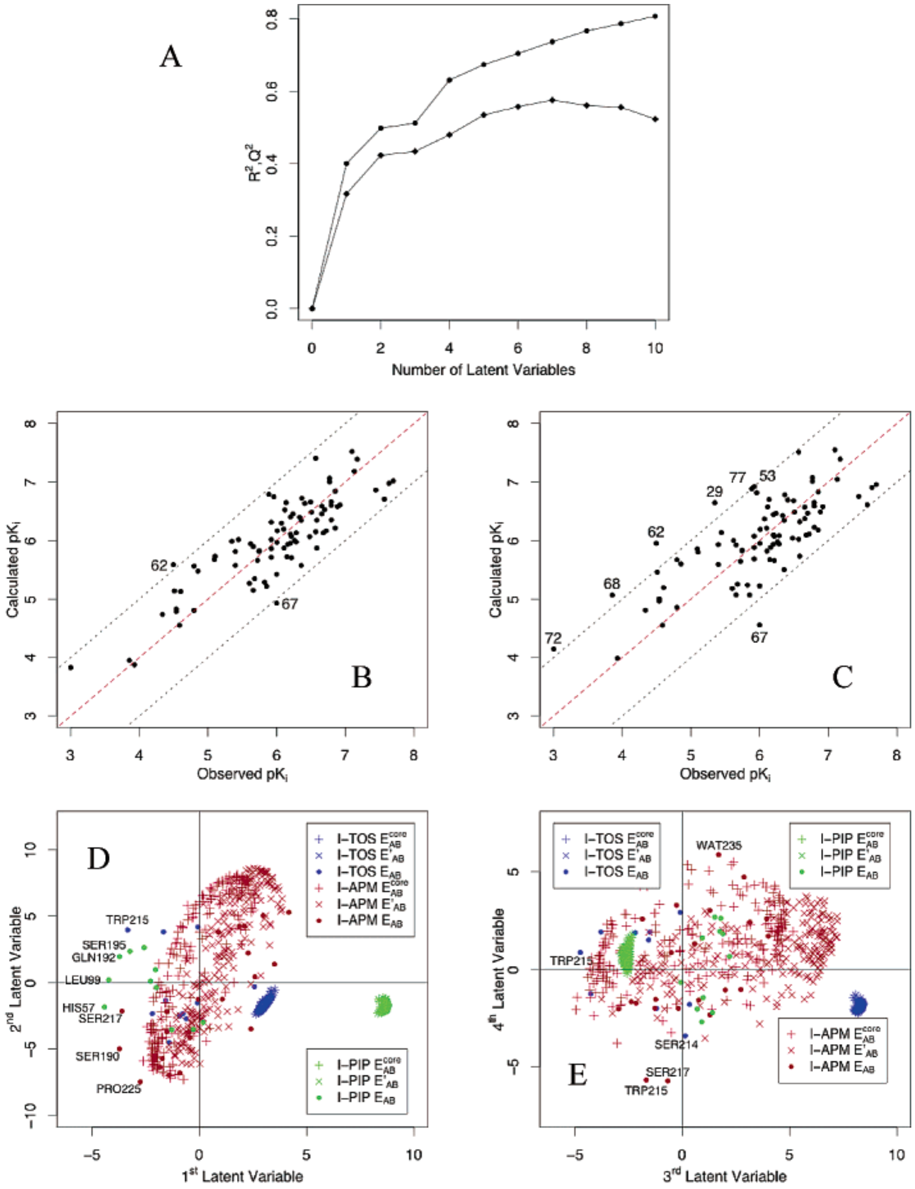


Figure 14. Model Lig3C PLS results. (A) R^2 and Q^2 versus number of latent variables. R^2 is represented as solid circles and Q^2 as solid diamonds. (B) Observed versus calculated pK_i values from internal validation. The red line represents the optimal correlation, while the blue lines are a pK_i unit from the optimal line. (C) Observed versus calculated pK_i values from LOO cross-validation. (D) Loading plot of latent variable 1 versus 2. (E) Loading plot of latent variable 3 versus 4.

Similar to model Lig1C the complexes are differentiated with the R group size playing a large role. Very large R^1 substituents lie to the left of the plot and decrease in size from left to right. The R^2 group sizes decrease from top to bottom. In other words, PC 1 explains the variation in R^1 and PC 2 explains the dissimilarity in R^2. The corresponding loading plot is shown in Figure 10.C, and again similar clustering occurs as model Lig1C. However, a deeper understanding of the key interactions can be constructed, due to the fragmentation of the ligands. After pruning model Lig3B, 11 I-**TOS**, 11 I-**PIP**, and 23 I-**APM** E_{AB} descriptors remained where I denotes any amino acid residue of the protein. The interactions between Trp215, Gln192, and Gly216 and the fragment **TOS** or R^1 are shown in Figure 10.D as dominant contacts. The larger the group at the R^1 position results in a greater interaction with Trp215. His57-, Gln192-, Leu99-, and Ser195-PIP dominates PC 2. Ser195 is a part of the catalytic triad where Leu99 lies between the **S2** and **S3** pockets. Both **S2** and **S3** are large pockets, and the results confirm the optimization of the size and shape of R^1 and R^2. Both interactions between the R groups and the residues of each pocket are dominated by hydrophobic contacts. As a way to demonstrate this, the change in activity by the substitution of 4-methylpiperidide (**4**) by *N*-methyl (**63**) at the R^2 position results in a 2.25 log unit loss in activity.

Partial Least Squares: Model Lig1C. A PLS model is generated to maximally explain the variance in *X* that correlates with *Y*. The statistical quantities of the Lig1C PLS model are shown in Table 3 such as R^2, Q^2, SDEC, and SDEP and the externally validated SDEC and SDEP.

R^2 and Q^2 plots against the number of latent variables (LVs) are shown in Figure 11.A. The R^2 values gradually increase with every additional latent variable as expected; however, the Q^2 value reaches a peak at 3 and tails off.

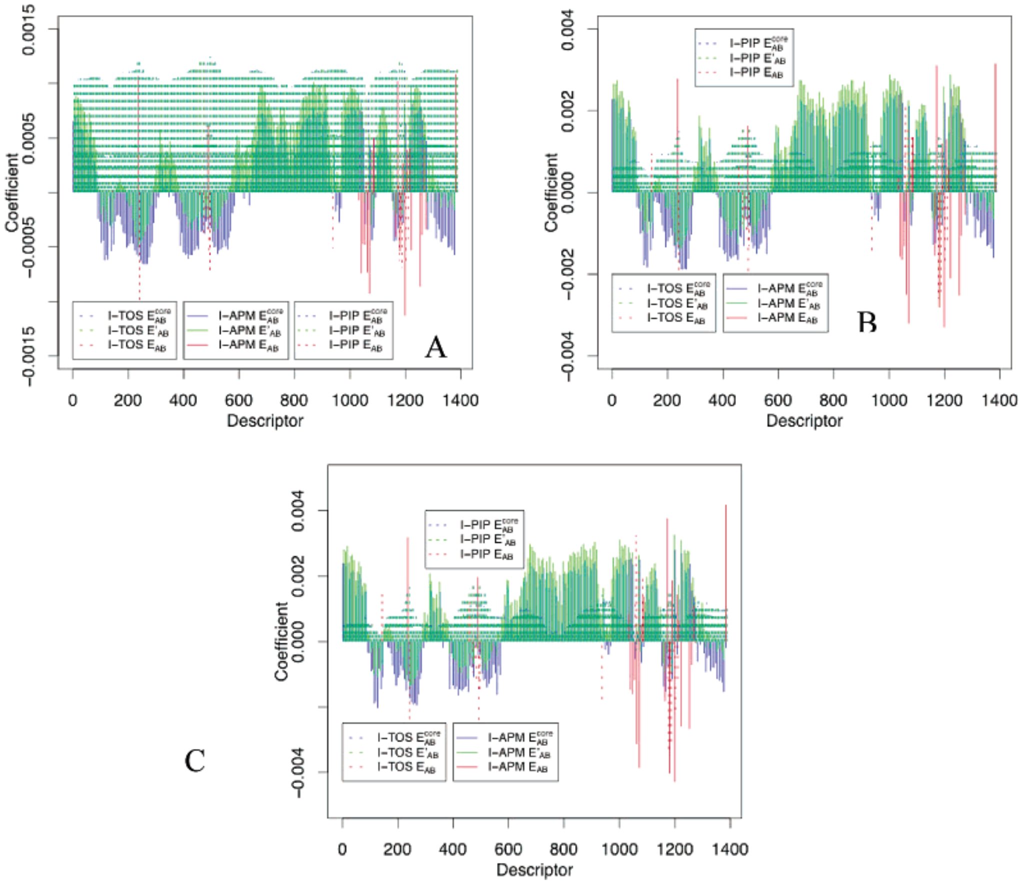


Figure 15. Model Lig3C PLS coefficient plots. (A) Latent variable 1. (B) Latent variable 2. (C) Latent variable 3.

Therefore, the optimal dimensionality of the PLS model involved 3 latent variables. Values higher than 0.5 for R^2 were considered statistically significant. Values greater than 0.4 for Q^2 were viewed as significant, and the optimal relationship between R^2 and Q^2 is the case where $R^2/Q^2 = 1$.[43] Although model Lig1C does not satisfy this equality, the 3 LV model explains 95% of the X matrix and 73% of the Y vector with a Q^2 of 0.59 and an SDEC of 0.45. The externally validated SDEP value of 0.42 is similar to that of the internal validation value, 0.56, and suggests a robust model. The predicted pK_i values are plotted against the experimental values in Figures 11.B, and the values predicted in the LOO cross-validation are shown in Figure 11.C.

When all the complexes were used to build the model, ligands **62** and **77** were presented as outliers (residual pK_i greater than 1 pK_i unit). Compounds **48** and **77** are similar structures where a proton is replaced with a methyl group. Model Lig1C predicts **48** to have a pK_i of 6.511 (6.658) with a residual of 0.147. However, the predicted value of **77** is 6.971 (5.886), a residual of 1.085. The methyl group in **77** is pointing directly into solution, while **48** has the ability to form hydrogen bonds with surrounding water molecules. The inclusion of solvation effects in our modeling would likely improve our ability to accurately model **77**. Compound **62** is also overestimated with a residual of 0.995. Nonetheless, it is considered more important that the predicted pK_i of similar structures **5** (6.328) and **17** (6.544) follows the same trend as that of the experimental values.

Using the LOO cross-validated approach the predicted values **29**, **67**, and **71** can also be identified as outliers. Compounds **29** and **71** are closely related to the high affinity ligand **10**. The order of the predicted pK_i's is incorrect when compared to the experimental values. Compounds **10**, **29**, and **71** are predicted to have activities of 6.514, 6.955, and 5.873, respectively. Compounds **10** and **29** differ in the orientation of the carboxylate group where Ser195 forms a hydrogen bond with **10**, while Gln192 interacts with **29**. The overestimation of the pK_i for **29** is probably due to the strong interaction with Gln192 that was "unseen" in the training phase. The distribution of the descriptors in the PLS model is shown in Figure 11.D where LV 1 is predominantly defined by the E_{AB}^{core} and E'_{AB} descriptors and the Trp215 E_{AB} term. Asp194, Pro225, and Gly196 define the second LV. His57 and Leu99 make strong contributions to both LVs.

These dominant interactions can be verified by looking at the PLS coefficient plots in Figure 12 and the Intermolecular Interaction Map (IMM) in Figure 13. The IMM plot highlights the dominant E_{AB} terms between the ligands and the receptor. IMMs allow the medicinal chemist to graphically represent the change in ligand fragment substitutions with an associated change in intermolecular energy at the residue level. For example, the interaction of **77** and **48** with His57 can easily be distinguished. Also the interaction of anthraquinone-2-sulfonyl (ACS) of **14** interacting with Thr98 is highlighted. The optimization of the ACS-like fragment with Thr98 may lead to stronger binding inhibitors. The score plots of the first and second latent variable are shown in Figure 11 (E and F). The compounds in the upper right quadrant are of high affinity, while those in the lower left

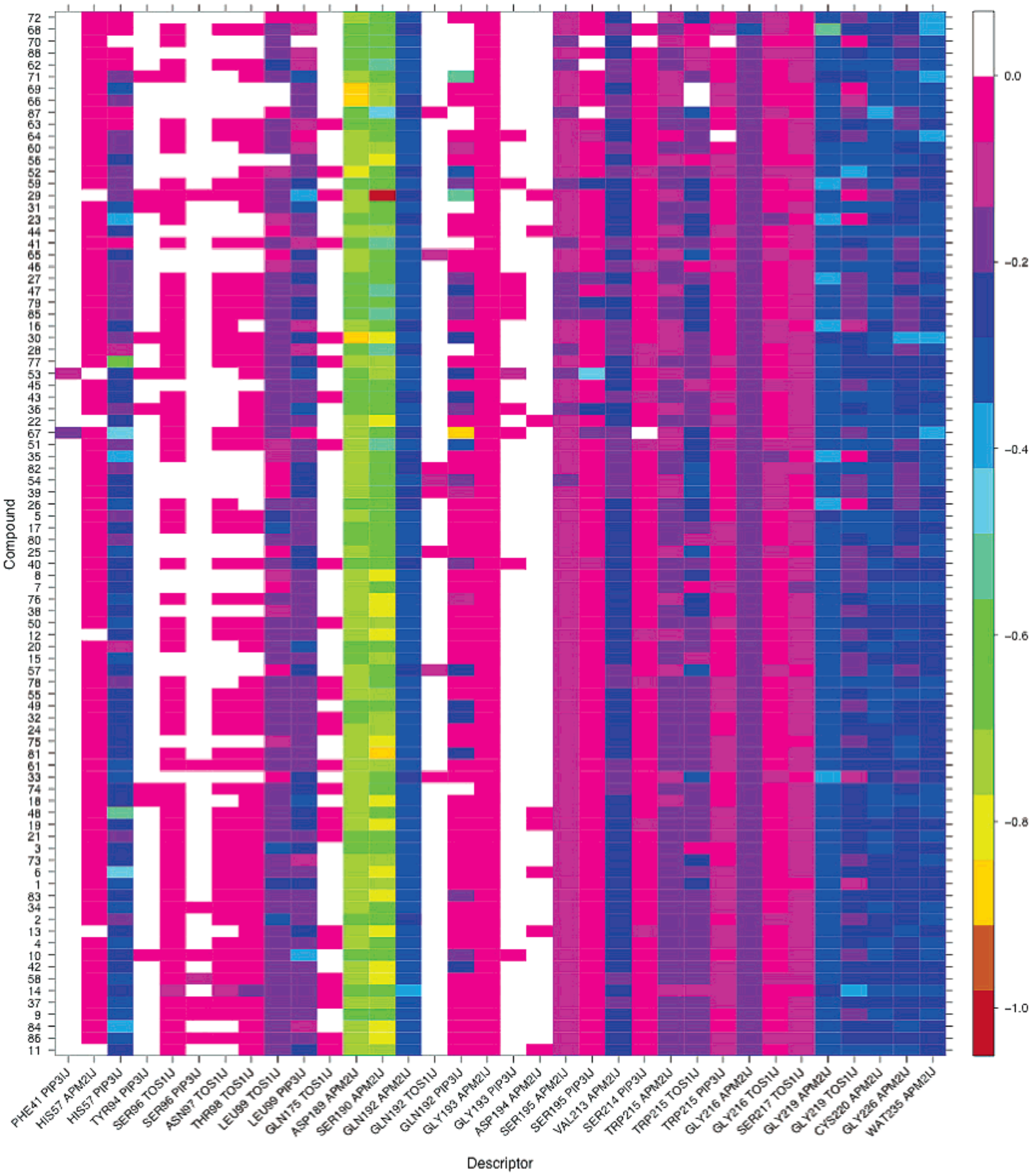


Figure 16. Model Lig3C intermolecular interaction map (IMM) of the important E_{AB} descriptors. The key residues of trypsin that interact with the triple fragment ligand (**APM**, **TOS**, and **PIP**) label the *x*-axis. The compounds on the *y*-axis are ordered with respect to activity. The activity decreases from top to bottom. The legend indicates the magnitude of the unscaled descriptor in eV.

quadrant have the lowest affinity.[44] The compounds incorrectly explained by a latent variable lie on the off-diagonal. Overestimated compounds populate the lower right quadrant, and underestimated compounds occupy the upper left quadrant. The first two LVs account for 94.18% of the variance in descriptor space. There is a greater spread around the optimal line in the score plot of the first LV than the second. The tight binding compounds **11** and **9** are correctly placed in the upper right quadrant of the score plot for LV 2, while the weaker binders such as **72** are also properly placed in the lower left quadrant. This suggests the importance of the E_{AB} descriptors even though they are in the minority. In the case of compound **11**, the electronegative oxygen atoms of the methyl ester on the piperazine group are pointed into **S2**, which is a favorable interaction. The 4-phenyl piperazide group of compound **84** fills the **S2** pocket and is predicted to have a high affinity. Compounds **10** and **71** differ by the stereochemistry of their carboxylate groups. **10** is a high affinity inhibitor, while **71** is a poor inhibitor. **71** is overestimated in the model (see Figure 11.C,F), and the analysis of the score plot confirms this.

Partial Least Squares: Model Lig3C. The statistical results for model Lig3C are shown in Table 3. The development of R^2 and Q^2 can be seen in Figure 14.A. The 7 LV model has an R^2 of 0.74, a Q^2 of 0.58, an SDEC of 0.45, an SDEP of 0.57. This model explains 96% of the X matrix and 74% of the activity variation. The external validated SDEP of 0.51 is similar to that of the internal LOO validated value thus suggesting a robust and predictive model. Statistically Lig1C and Lig3C are very similar; however, in terms of interpretability Lig3C is far superior. Lig3C allows the modeler to assess the interaction energy change upon

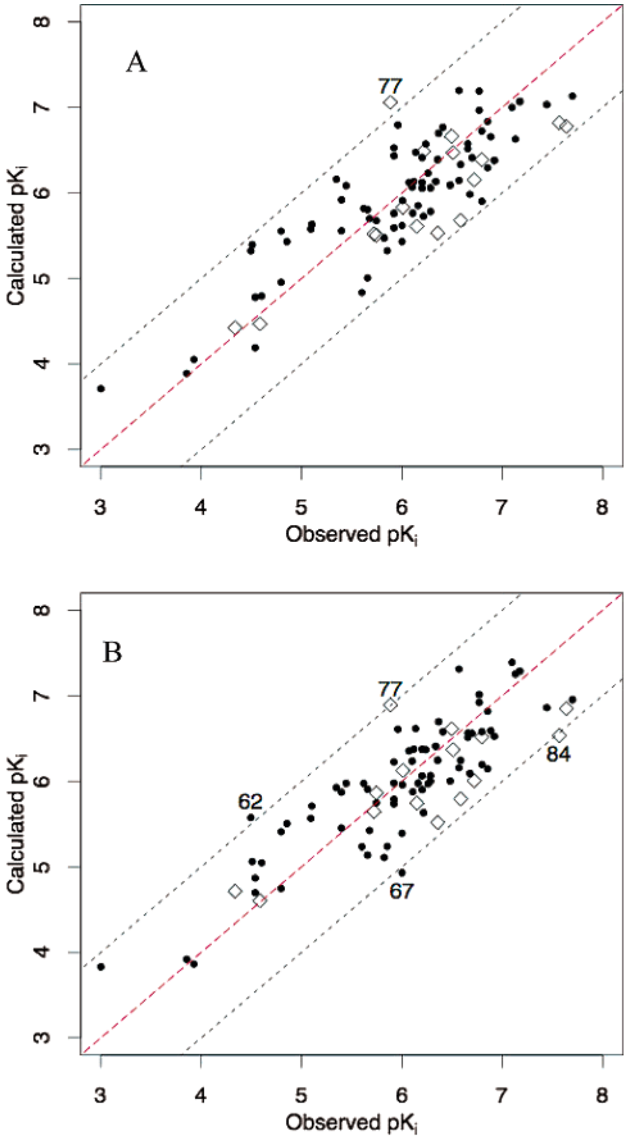


Figure 17. Models Lig1C and Lig3C derived using the training (circles) and prediction (diamonds) sets outlined by Böhm et al. (A) Model Lig1C. (B) Model Lig3C. The red line represents the optimal correlation, while the blue lines are a pK_i unit from the optimal line.

fragment substitution. The predicted pK_i values are plotted against the experimental values in Figure 14.B, and the values predicted by LOO cross-validation are shown in Figure 14.C. The compounds **62** and **67** are outliers in the model. Compounds **29**, **53**, **68**, **72**, and **84** are also outliers in the LOO scheme. The loading plot for LV 1 versus LV 2 and LV3 versus LV 4 are shown in Figure 14D,E. The interactions of Ser195, Gln192, Leu99, and His57 with the fragment **PIP** dominates the first LV, while the interactions of Ser190 and Pro225 with the fragment **APM** dictates the second LV. LV 3 is characterized by the Trp215 interaction with fragment **TOS**, and LV 4 is distinguished by the interaction of Wat235, Ser217, and Trp215 with **APM**. The coefficient plots of the first three latent variables are shown in Figure 15, which complement the loading plots. Similar to the Lig1C model an IMM can help to decipher the reasons for outliers and low affinity of certain inhibitors. The IMM of the important E_{AB} descriptors are shown in Figure 16. The variance in activity of the inhibitors can be predominantly explained by the interactions with the amino acid residues listed above. Interestingly, Asp189 is not significant in the model; however, as shown in the IMM, the interactions with Asp189 are strong and are essential for binding.

Compound **62** presents as an outlier in both SE-COMBINE models. Ligand **62** is overestimated in this model with a residual value of 1.092 pK_i units. Similar to model Lig1C the trend of predicted activity values of **5**, **17**, and **62** are in the same order as experiment. The predicted value for **77** is overestimated compared to **48**, which is a similar result to model Lig1C in both training and cross-validation. The IMM highlights the difference between the two by considering the interaction between His57 and **PIP**. The interaction is stronger for **77** compared to **48** and results in overbinding. The interactions of Gln192 and **PIP** of compounds **71**, **29**, and **67** are stronger compared to the rest of the compounds as shown in the IMM. The activity of **29** is overestimated compared to compounds **10** and **71** due to the strong interaction with Gln192. **53** and **67** interact with Phe41 deep in the **S2** pocket; however, this does not translate into increased activity. The activity values are concentrated between 5.5 and 7 pK_i units, and so the modeling of compound **72** is challenging due to its low activity. **72** is a leverage point, meaning that it has a high influence on the model. Experimentally the addition of a carboxylate group to **78** creating **60** causes a pK_i drop of 1.561. The IMM suggests that the carboxylate interacts favorably with Gln192; however, poorer binding activity results.

Model Lig1C and Lig3C Derived using Predefined Training and Prediction Sets. To test the ability of SE-COMBINE to be used in a real lead optimization situation QSAR models were built using the training and prediction sets outlined by Böhm et al. where the first 72 compounds were placed in the training set and the remaining 16 compounds made up the prediction set. Two different studies have been carried out using this division of the data set. Böhm et al.[35] reported a CoMFA and CoMSIA study in 1999 and Robert et al.[37] described a quantum similarity study in 2000. Again, the Lig1C and Lig3C descriptors were used to construct two QSAR models, and the results of these models are shown in Table 4. The plots of observed versus predicted activity values of both the training and prediction sets for both models are shown in Figure 17. The predicted pK_i residual of compound **77** in model Lig1C is greater than one pK_i unit, while **62**, **67**, and **84** have similar residuals in model Lig3C. A comparison was then made between the different methods used to predict the binding affinity of the 88 compounds to trypsin. On initial inspection the SE-COMBINE methods seems to underperform its nonreceptor based counterparts as shown in Table 5. The CoMSIA approach has the highest Q^2 value of 0.75 and the lowest SDEP. The CoMSIA approach calculates the steric occupancy, partial atomic charges, local hydrophobicity, and hydrogen-bond donor and acceptor properties at each grid point. Currently, SE-COMBINE only considers the electrostatic interaction between the receptor and ligand, and such effects as receptor and ligand desolvation and dispersion are neglected. On the basis of this, it is not an unexpected result that SE-COMBINE does not perform as well as these methods. The Fragment QS-SM method removed three compounds from the training set; however, the Q^2 is still only 0.69, and it has an SDEP of 0.51 which is similar to the SE-COMBINE models. The QSM methods have a high number of LVs compared to the number of descriptors, and overfitting could be an issue as well.

Conclusion

This research describes the derivation and implementation of a new receptor-based QSAR called SE-COMBINE. The validation of SE-COMBINE was used as an investigation of the interactions between trypsin and a series of inhibitors. The interactions between key residues of the protein and fragments of the ligand were elucidated, and their changes were compared to experimental data. The research shows that SE-COMBINE can be used in a lead optimization scheme in structure-based drug-design. This method allows the chemist to investigate the gain or loss of interaction energy upon fragment substitution. SE-COMBINE was not directly compared to another receptor-based method as a control experiment; however, considering that SE-COMBINE includes effects such as charge transfer and polarization, it is possible that it would outperform its molecular mechanics counterparts, such as COMBINE and MM/PBSA. The models generated using SE-COMBINE were competitive when compared to nonreceptor based QSARs, considering that it uses an incomplete energy function.

The decomposed interaction energies have shed light onto the key interactions between trypsin and a series of benzamidine-based inhibitors. Using current statistical and graphical tools PWD and SE-COMBINE has the potential to be a powerful technique in structure-based drug design. From the detailed analysis of the protein−ligand interactions, predictions of new and improved inhibitors can be made. Model Lig3C highlighted the important ligand fragment−protein residue interactions and thus allows a computational chemist to create hypothetical virtual compounds and predict their activity using the model. For example, the ACS group of **14** could be optimized to enhance the interactions with Thr98 or modify other scaffolds to include this interaction.

SE-COMBINE is not limited to drug design. Problems such as protein-decoy-discrimination, protein stability, and protein metal ion selectivity and in silico protein mutagenesis studies can be targeted using this approach. Investigating such problems with SE-COMBINE can only lead to a better understanding of molecular stability, recognition, selectivity, and ultimately a more complete understanding of molecular interactions.

SE-COMBINE in its current form does not decompose a complete energy function. Recent developments have shown that the solvation free energy of binding can be partitioned using either a Generalized-Born or Poisson−Boltzmann approach. Dispersion effects are neglected in QM methods. These could easily be included in the form of a Lennard-Jones ($(1/R^6)$) potential. The attractive part of the LJ potential lends itself to be partitioned; in fact it is widely used in MM potentials. The entropy term of the master binding equation thus remains. Entropy is not an intermolecular interaction, and so it is almost impossible to have a complete pairwise function. However, it could easily be added to the potential function without being partitioned. Hence, future work will add solvation, dispersion, and entropy components to SE-COMBINE which, in effect, would result in a pairwise QMSCORE.[25]

Acknowledgment. The authors are grateful to Gerhard Klebe for providing the trypsin data set and to Rajarshi Guha for many useful discussions. We thank the NSF (MCB-0211639) and the NIH (GM 44974) for financial support of this research. The helpful comments by the anonymous reviewers are also acknowledged.

References

(1) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47* (12), 3032−3047.

(2) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

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

(4) Wade, R. C.; Ortiz, A. R.; Gago, F., Comparative binding energy analysis. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 19−34.

(5) Ortiz, a. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of Drug-Binding Affinities by Comparative Binding-Energy Analysis. *J. Med. Chem.* **1995**, *38* (14), 2681−2691.

(6) Kuhn, B.; Kollman, P. A. Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of their Relative Affinites by Combination of Molecular Mechanics and Continuum Solvation Models. *J. Med. Chem.* **2000**, *43* (20), 3786−3791.

(7) Wang, W.; Lim, W. A.; Jakalian, A.; Wang, J.; Wang, J.; Luo, R.; Bayly, C. I.; Kollman, P. A., An Analysis of the Interactions between the Sem-5 SH3 Domain and its Ligands Using Molecular Dynamics, Free Energy Calculations, and Sequence Analysis. *J. Am. Chem. Soc.* **2001**, *123*, 3986−3994.

(8) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 131−150.

(9) Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109−130.

(10) Pople, J. A.; Beveridge, D. L. In *Approximate Molecular Orbital Theory*; McGraw-Hill: New York, 1970.

(11) van der Vaart, A.; Gogonea, V.; Dixon, S. L.; Merz, K. M., Jr. Linear Scaling Molecular Orbital Calculations of Biological Systems Using the Semiempirical Divide and Conquer Method. *J. Comput. Chem.* **2000**, *21*, 1494−1504.

(12) Dixon, S. L.; Merz, K. M., Jr. Fast, Accurate Semiempirical Molecular Orbital Calculations for Macromolecules. *J. Chem. Phys.* **1997**, *107*, 879−893.

(13) Dixon, S. L.; Merz, K. M., Jr. Semiempirical Molecular Orbital Calculations with Linear System Size Scaling. *J. Chem. Phys.* **1996**, *104*, 6643−6649.

(14) Yang, W.; Lee, T.-S. A Density-matrix Divide-and-conquer Approach for Electronic Structure Calculations of Large Molecules. *J. Chem. Phys.* **1995**, *103* (13), 5674−5678.

(15) Li, X. P.; Nunes, R. W.; Vanderbilt, D. Density-Matrix Electronic-Structure Method with Linear System-Size Scaling. *Phys. Rev. B* **1993**, *47* (16), 10891−10894.

(16) Stewart, J. J. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *Int. J. Quantum Chem.* **1996**, *58* (2), 133−146.

(17) Dixon, S. L.; van der Vaart, A.; Gogonea, V.; Vincent, J. J.; Brothers, E. N.; Suárez, D.; Westerhoff, L. M.; Merz, K. M., Jr. *DIVCON99*, The Pennsylvania State University: 1999.

(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(19) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209−220.

(20) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.

(21) Thiel, W.; Voityuk, A. A. Extension of MNDO to d Orbitals: Parameters and Results for the Second-Row Elements and for the Zinc Group. *J. Phys. Chem.* **1996**, *100*, 616−626.

(22) Dewar, M. J. S.; Thiel, W. Ground States of Molecules. 38. The MNDO method. Approximations and Parameters. *J. Am. Chem. Soc.* **1977**, *99* (15), 4899−4907.

(23) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. PDDG/PM3 and PDDG/MNDO: Improved semiempirical methods. *J. Comput. Chem.* **2002**, *23* (16), 1601−1622.

(24) Raha, K.; Merz, K. M., Jr. A Quantum Mechanics Based Scoring Function: Study of Zinc-ion Mediated Ligand Binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020−1021.

(25) Raha, K.; Merz, K. M., Jr. Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein−ligand complexes. *J. Med. Chem.* **2005**, *48* (14), 4558−75.

(26) Raha, K.; van der Vaart, A. J.; Riley, K. E.; Peters, M. B.; Westerhoff, L. M.; Kim, H.; Merz, K. M., Jr. Pairwise Decomposition of Residue Interaction Energies Using Semiempirical Quantum Mechanical Methods in Studies of Protein−Ligand Interaction. *J. Am. Chem. Soc.* **2005**, *127* (18), 6583−6594.

(27) Wang, B.; Raha, K.; Merz, K. M., Jr. Pose Scoring by NMR. *J. Am. Chem. Soc.* **2004**, *126* (37), 11430−11431.

(28) Davie, E. W.; Fujikawa, K.; Kisiel, W. The Coagulation Cascade - Initiation, Maintenance, and Regulation. *Biochemistry* **1991**, *30* (43), 10363−10370.

(29) Silverman, R. B. *The organic chemistry of enzyme-catalyzed reactions*; Academic Press: San Diego, 2002.

(30) Turk, D.; Sturzebecher, J.; Bode, W. Geometry of Binding of the N-Alpha-Tosylated Piperidides of Meta-Amidino-Phenylalanine, Para-Amidino-Phenylalanine and Para-Guanidino-Phenylalanine to Thrombin and Trypsin - X-ray Crystal-Structures of Their Trypsin Complexes and Modeling of Their Thrombin Complexes. *FEBS Lett.* **1991**, *287* (1−2), 133−138.

(31) 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* (3), 593−614.

(32) Mares-Guia, M.; Shaw, E. Studies on the active center of trypsin; the binding of amidines and guanidines as models of the substrate side chain. *J. Biol. Chem.* **1965**, *240*, 1579−1585.

(33) Sturzebecher, J.; Prasa, D.; Hauptmann, J.; Vieweg, H.; Wilkstrom, P. Synthesis and structure−activity relationships of potent thrombin inhibitors: Piperazides of 3-amidinophenylalanine. *J. Med. Chem.* **1997**, *40* (19), 3091−3099.

(34) Sturzebecher, J.; Prasa, D.; Wikstrom, P.; Vieweg, H. Structure−Activity-Relationships of Inhibitors Derived from 3-Amidinophenylalanine. *J. Enzymol. Inhib.* **1995**, *9* (1), 87−99.

(35) Böhm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure−activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42* (3), 458−477.

(36) Klebe, G. Comparative Molecular Similarity Indices: CoMSIA. In *3D QSAR in Drug Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic Publishers: Great Britain, 1998; Vol. 3, p 87.

(37) Robert, D.; Amat, L.; Carbo-Dorca, R. Quantum similarity QSAR: Study of inhibitors binding to thrombin, trypsin and factor Xa, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem.* **2000**, *80* (3), 265−282.

(38) Murcia, M.; Ortiz, A. R. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem.* **2004**, *47* (4), 805−820.

(39) Cornell, W. D.; Cieplak, P.; Baylay, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field For the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(40) R *R: A Language and Environment for Statistical Computing, 2.0.1*; R Development Core Team: R Foundation for Statistical Computing: Vienna, Austria, 2005.

(41) Mevik, B. H.; Cederkvist, H. R. Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *J. Chemom.* **2004**, *18* (9), 422−429.

(42) Cundari, T. R.; Sarbu, C.; Pop, H. F. Robust fuzzy principal component analysis (FPCA). A comparative study concerning interaction of carbon−hydrogen bonds with molybdenum-oxo bonds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1363−1369.

(43) Golbraikh, A.; Tropsha, a., Beware of q(2)! *J. Mol. Graphics* **2002**, *20,* (4), 269−276.

(44) Stanton, D. T. On the physical interpretation of QSAR models. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (5), 1423−1433.

CT050284J