EMBM − A New Enzyme Mechanism-Based Method for Rational Design of Chemical Sites of Covalent Inhibitors

Tamar Traube, Subramaniam Vijayakumar, Michal Hirsch, Neta Uritsky, Michael Shokhen,* and Amnon Albeck*

The Julius Spokojny Bioorganic Chemistry Laboratory, Department of Chemistry, Bar Ilan University, Ramat Gan 52900, Israel

Received August 26, 2010

We introduce an enzyme mechanism-based method (EMBM) aimed at rational design of chemical sites (CS) of reaction coordinate analog inhibitors. The energy of valence reorganization of CS, caused by the formation of the enzyme−inhibitor covalent complex, is accounted for by new covalent descriptors W1 and W2. We considered CS fragments with a carbonyl reactivity center, like in native protease substrates. The W1 and W2 descriptors are calculated quantum mechanically on small molecular clusters simulating the reaction core of the formed covalent tetrahedral complex, anionic TC(O−) or neutral TC(OH). The modeling on a reaction core allows generation of various CS and corresponding TC(O−) and TC(OH) as universal building blocks of real inhibitors and their covalent complexes with serine or cysteine hydrolases. Moreover, the approach avoids the need for 3D structure of the target enzyme, so EMBM may be used for ligand-based design. We have built a chemical site of inhibitors (CSI) databank with pairs of W1 and W2 descriptors precalculated for both $CH_3O(-)$ and $CH_3S(-)$ nucleophiles for every collected CS fragment. We demonstrated that contribution of a CS fragment to the binding affinity of an inhibitor depends on both its covalent reorganization during the chemical transformation and its noncovalent interactions in the enzyme active site. Consequently, prediction of inhibitors binding trend can be done only by accounting for all of these factors, using W1 and W2 in combination with noncovalent QSAR descriptors.

INTRODUCTION

Small molecule enzyme inhibitors satisfying the Lipinski rule of five[1] are frequently used as candidates in the design of new drugs.[2] Most methodologies of computer-assisted drug design (CADD) focus on the enzyme−inhibitor recognition by noncovalent interactions.[3] The binding affinity of these interactions correlates with the enzyme−inhibitor contact surface area.[4] The mean values of experimental binding affinities, $\lg K_a$, are 3.7 for enzyme−substrate, 8.6 for enzyme−noncovalent inhibitor, and 16 for enzyme-transition state complex.[4] The latter was derived from Wolfenden's "catalytic proficiency", $1/K_{tx} = (k_{cat}/K_m)/k_{uncat}$, the ratio between the rate constants of enzyme catalyzed process and the uncatalyzed reaction in water.[5] Stabilization of the transition state (TS) of an enzymatic reaction is exploited for the design of transition-state analog (TSA) inhibitors,[6] some of which have already been developed into commercial drugs.[7]

There are two types of TSA inhibitors: noncovalent and reversible covalent. The former mimics the shape of TS and utilizes electrostatic and H-bond interactions with the pre-organized polar groups in the enzyme active site,[8] analogous to those of the TS.[6] The latter undergoes a catalytic reversible chemical reaction, similar to the native reaction with the substrate, to form a thermodynamically stable complex which resembles the catalytic reaction TS. The α-ketoheterocyclic inhibitors of serine and cysteine proteases represent the latter family of TSA inhibitors.[9] To differentiate between the two types of TSA inhibitors, Lipscomb has introduced the term "reaction-coordinate analogue", RCA, for covalent TSA inhibitors.[10]

RCA inhibitors can be considered as formally comprised of two parts: the chemical site (CS), responsible for the covalent binding, and the recognition site (RS), dominating the selectivity of the inhibitor toward the target enzyme.[11,12] We have previously introduced an original method that allows a separate design of CS fragments of RCA inhibitors.[12] The method uses a series of isoselective RCA inhibitors (with identical RS and different CS fragments), in which the relative binding affinity depends on their CS fragments only. The method explicitly accounts only for the covalent interactions between the inhibitor CS fragment and the attacking nucleophile of the enzyme.[12] Recently we have introduced new and improved covalent QSAR descriptors W1 and W2.[13] We demonstrated that the W1 and W2 descriptors, in combination with quantitative structure−activity relationship (QSAR) analysis of the binding trend of RCA inhibitors with varied CS fragments, can serve as a mechanistic tool, differentiating between the catalytic mechanisms of serine and cysteine proteases.

In the present study, we introduce a new QSAR method for analysis of CS interactions in the enzyme active site. It is unique by accounting for both covalent, by W1 and W2, and noncovalent interactions, by conventional 2D descriptors implemented in most drug design software packages. It provides a practical tool for separate rational design of CS fragments of RCA inhibitors, as demonstrated for proteolytic

* Corresponding authors. E-mail: shokhen@mail.biu.ac.il (M.S.), albecka@mail.biu.ac.il (A.A.).

10.1021/ci100330y © 2010 American Chemical Society
Published on Web 11/19/2010

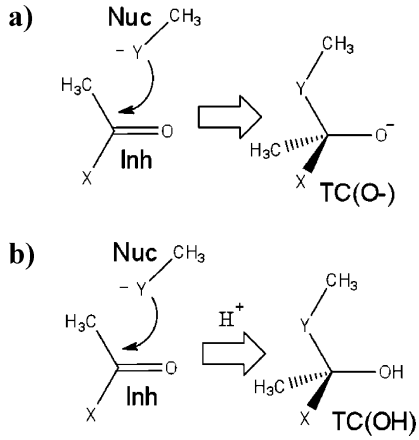

Figure 1. The first catalytic step of proteases, simulated by molecular clusters of the enzyme Nuc−Inh reaction core. Formation of (a) an anionic TC(O−) and (b) a neutral TC(OH). Y = O or S, in serine or cysteine nucleophile, correspondingly. X is the varied substituent modifying the reactivity of the carbonyl group of the ligand.

enzymes. Whereas other methods can be classified as either structure-based or pharmacophore/ligand-based methods, our method should best be classified as enzyme mechanism-based method (EMBM).

METHODS

Covalent Descriptors. EMBM is aimed at the design of CS fragments of RCA inhibitors. The method uses series of isoselective RCA inhibitors, the relative binding affinities of which depend on the variation of their CS fragments only. The binding affinity trend is predicted on small molecular clusters representing the enzyme−inhibitor reaction core in the active site.[12,13] The clusters are composed from the functional groups simulating the enzyme nucleophile (Nuc) and the CS fragment of the inhibitor (Inh) (Figure 1). In all considered series, the electrophilic center of the inhibitor is a carbonyl group modified by different substituents X, such as α-ketoheterocycles.[9] The nucleophilic Y is either an O or a S atom, corresponding to the Ser or Cys catalytic residues. We considered two variants of the covalent tetrahedral complexes, either anionic TC(O−) (Figure 1a) or neutral protonated species TC(OH) (Figure 1b).

The energy of the new enzyme−inhibitor covalent bond formation (O−C or S−C), partial cleavage of the former carbonyl bond, and protonation of the former carbonyl oxygen atom in the case of TC(OH) are accounted for by W1 and W2, the reaction enthalpy values, calculated by eqs 1 and 2:[13]

$$W1 = H_f[TC(O-)] - H_f(Inh) - H_f(Nuc) \quad (1)$$

$$W2 = H_f[TC(OH)] - H_f(Inh) - H_f(Nuc) \quad (2)$$

Molecular structures of every Inh, TC(O−), and TC(OH) were subjected to conformational analysis by the BEST algorithm[14] in frames of cff force field[15] implemented in Accelrys Discovery Studio (DS).[16] The most stable conformer of every species was fully optimized in PM6 semiempirical quantum mechanical Hamiltonian,[17] applying MOPAC2009.[18] W1 and W2 were calculated as heats of formation, H_f, of TC(O−) and TC(OH), respectively, at 25 °C in PM6 Hamiltonian by MOPAC2009. The heterogeneous combined protein/water environmental effect, considerably influencing the W1 and W2 values, was accounted for by the quantum mechanics/self-consistent reaction field (virtual solvent) [QM/SCRF(VS)] approach[19] by the continuum solvation COSMO algorithm[20] implemented in MOPAC2009. W1 and W2 were calculated by eqs 1 and 2 for every CS on a wide interval (from 1 to 80) of dielectric constant of the polarized continuum. Experimental pK_i or pIC50 were presented as a linear function of two independent variables W1 and W2. The values of W1 and W2 identified at the best correlation with max R^2 and min root-mean-square (rms) error in multiple linear regression analysis were used as descriptors in the following QSAR analysis of the binding trend of the target enzyme with RCA inhibitors. W1 and W2 of dozens of different CS's can be calculated this way on a desktop PC in a reasonable time and do not require any 3D structural information about the target enzyme. We have precalculated W1 and W2 indices for about 300 different CS groups forming covalent TC's with two types of nucleophiles CH_3O^- and CH_3S^- corresponding to serine and cysteine hydrolases. The results are accumulated in our chemical site of inhibitors (CSI) databank.

By definition, W1 accounts for the modulation of the electrophilicity of the carbonyl by the variation of CS. The W2 index accounts for two energetic effects: formation of the enzyme−inhibitor covalent bond and PA of the carbonyl oxygen or pK_a of TC(OH). We introduce W1 and W2 as new covalent QSAR descriptors in combination with well established 2D noncovalent and topological descriptors implemented in the DS package,[16] for the design of CS fragments of RCA inhibitors. The QSAR models were generated and optimized by genetic function approximation (GFA),[21] implemented in DS. GFA selects the most relevant descriptors dominating the inhibitors binding trend in QSAR models. We used different formulations of input parameters controlling the GFA algorithm in DS software,[16] such as linear or quadratic models, as well as a different maximal number of terms in the final regression equation. Condition of max R^2 and min rms error was used for the selection of the QSAR models generated by the GFA algorithm. We gave preference to models with minimal independent variables between those with close R^2 and rms error on the training set. We also used a multiple linear regression analysis protocol, also implemented in the DS, as an alternative for the QSAR models optimization. In the latter case, we analyzed different combinations of QSAR descriptors that most frequently appeared in the best QSAR models generated by the GFA algorithm as well as some other descriptors based on our knowledge and chemical intuition. Finally, the best QSAR models were examined on the test sets. The best QSAR models of all considered enzymes are presented in Tables S1−S8 in the Supporting Information. All 2D descriptors used as the input data together with the W1 and W2 descriptors in the GFA algorithm for the generation of the QSAR models are presented in Table S9.

Validation of predictability of a QSAR model is a crucial step. It was demonstrated that the widely accepted leave-one-out (LOO) test is inadequate in many cases; no correlation was found between the values of cross-validated q^2 for the training set and the predictive ability for the test set.[22] Thus, the only way to establish reliability of a QSAR model is by its validation on a test set. Therefore, we examined

every QSAR model by prediction on test sets, provided we had sufficient number of experimental points for separation to training and test sets.

CSI Databank. We built a CSI databank into which we collect and update literature molecular and structural data of isoselective inhibitors of serine and cysteine hydrolases and the corresponding experimental binding constants, K_i and IC50 values. We have already accumulated about 300 precalculated different CS fragments. Calculations were conducted by the EMBM protocol. The databank contains the following information: experimental IC50 and inhibitors binding constants, K_i, with references to the source articles; molecular structures in 2D and 3D presentations of the CS fragments of each inhibitor; and heats of formation, H_f, for every CS calculated at 25 °C in water ($\varepsilon = 78.4$); molecular structures in 2D and 3D presentations of TC(O−) and TC(OH) formed by Ser or Cys nucleophiles (Figure 1); heats of formation, H_f, of every TC(O−) and TC(OH) calculated by eqs 1 and 2 at 25 °C in a wide interval (1−80) of dielectric constants, ε; and W1 and W2 descriptors. The data in the CSI bank are saved in standard data file (SDF) format, so they can be opened and used by most molecular modeling software packages. The CSI databank is continuously updated with new CS fragments. We used the databank for the current work. Figures 1S−4S in the Supporting Information present examples illustrating the information collected in the CSI databank.

RESULTS AND DISCUSSION

We examined the method validation on different series of isoselective RCA inhibitors targeted at enzymes with published K_i and IC50 values: the serine proteases thrombin,[23] trypsin,[23] and human neutrophil elastase (HNE);[24] the serine hydrolase fatty acid amide hydrolase (FAAH);[25] and the cysteine proteases cathepsin K,[26] HRV 3C,[27] and caspase 3.[28] In the present work, the reaction core of CS is a carbonyl group substituted by a varying X (Inh in Figure 1). If variation of X considerably changes the electron distribution on the reactivity center (the carbonyl group), then the covalent W1 and W2 descriptors play a key role in the binding trend of RCA inhibitors. W1 accounts for the modulation of electrophilicity of the carbonyl group by the varied X (eq 1) and dominates the inhibitors binding trend in formation of anionic TC(O−) for a strong Ser nucleophile. By definition (eq 2 and ref 13), the W2 index accounts for two energetic effects: formation of the enzyme−inhibitor covalent bond and PA/pK_a of TC(OH). The latter effect dominates the biding trend of RCA inhibitors with a weak Cys nucleophile.[13] Therefore, W2 plays an exclusive role in the QSAR models of cysteine proteases. Combination of covalent W1 or W2 with noncovalent 2D descriptors considerably improves the predictability of QSAR models. The physical background and detailed analysis of the different role of W1 and W2 descriptors in the QSAR models of Ser and Cys nucleophiles are presented in ref 13 (see also Supporting Information).

For the EMBM validation we split the initial set of compounds into training and test sets, such that the test set contains ~20% of the total number of compounds. The compounds for the test set were selected from the original set in such a way that their K_i or IC50 values were randomly

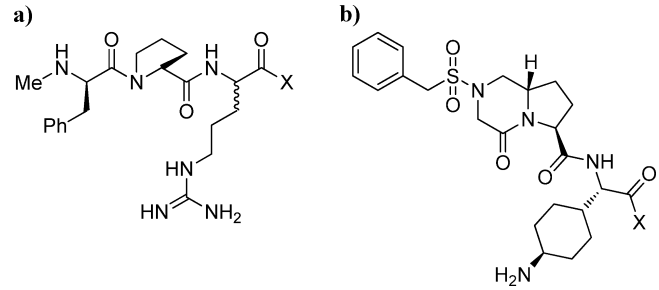

Figure 2. Peptidyl α-ketoheterocyclic inhibitors of thrombin and trypsin. X is a varied substituent at the carbonyl group in the CS of inhibitors: (a) training[23a] and (b) test[23b] sets.

but uniformly distributed in the range of the values of the whole set. The QSAR model generated on a training set was examined on its predictability on a test set. We analyzed three variants of QSAR models in order to identify the separate role of covalent and noncovalent descriptors in the predictability of the binding trend of varied CS's on the corresponding test sets: the total set of independent variables-covalent descriptor (W1 in serine proteases or W2 in cysteine proteases) and 2D noncovalent descriptors; covalent (W1 or W2) descriptor only; and noncovalent 2D descriptors. We have to stress that the number of published series of isoselective RCA inhibitors is very limited; the maximal number of varied CS rarely exceeds 25 in one series. In cases where the initial set of compounds was too small, so the test set cannot contain five or more compounds, we used only QSAR analysis on the full set of compounds, omitting the prediction analysis on a test set. The Supporting Information contains tables with detailed presentation of independent variables and regression coefficients of QSAR models generated for every considered enzyme.

Thrombin and Trypsin. We used K_i's experimental data for compounds in Table 5 in ref 23a as a training set (Figure 2a) and compounds from Table 1 in ref 23b as a test set (Figure 2b) for the serine proteases thrombin and trypsin.

The best prediction on the test set for thrombin, $R^2 = 0.814$, was obtained by the QSAR model containing the combined set of variables identified by GFA: W1 and four 2D descriptors (Figure 3a and Table S1 in the Supporting Information). We note that the training and the test sets of isoselective inhibitors have different recognition fragments, RS. This supports two basic characteristics of the EMBM approach: The varied CS dominates the binding trend in isoselective inhibitors, and CS of RCA inhibitors can be designed on a modeling set with one RS and then be applied to another set of inhibitors with a different RS optimal for the target enzyme. Sequential exclusion of variables demonstrated a key role of the W1 descriptor in the QSAR predictability, in contrast to the inadequate prediction of the binding trend based on noncovalent descriptors only (Figure 3b and 3c and Table S1 in the Supporting Information).

Using the same training set (Table 5 in ref 23a) and test set (Table 1 in ref 23b) for trypsin, we observed a very good correlation between calc pK_i vs exp pK_i, $R^2 = 0.887$, and poor prediction, $R^2 = 0.118$, in the QSAR model containing W1 and 4 noncovalent descriptors (Figure 4 and Table S2 in the Supporting Information). The good prediction for thrombin and poor for trypsin originates in the RS. Comparison of the two series of inhibitors (Table 5 in ref 23a and Table 1 in ref 23b) shows that the RS's in both series

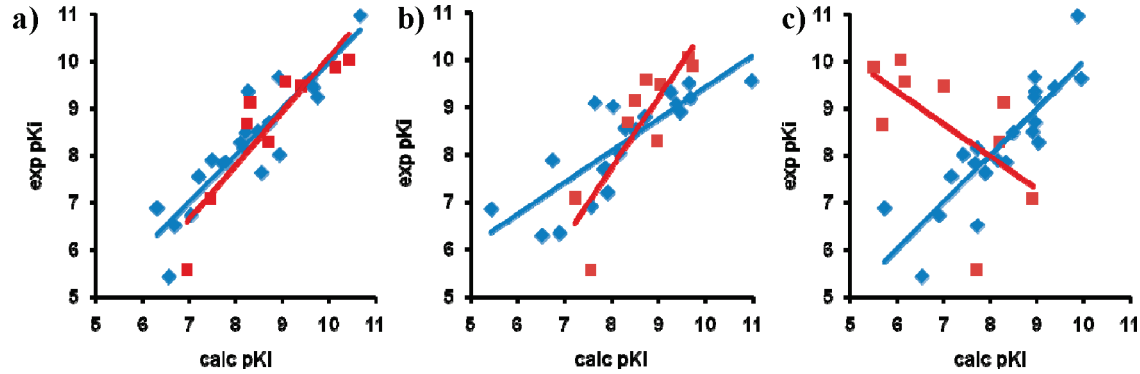

Figure 3. Correlations of experimental and calculated pK_i's in QSAR models for thrombin generated on the training set (blue) and examined on the test set (red). The training set contains 21 compounds: 1b, 4, and 50−68 from Table 5 in ref 23a. The test set consists of 9 compounds: 2b−f and 2h−k from Table 1 in ref 23b. (a) W1 and 2D descriptors, $y = 1x + 5 \times 10^{-7}$, $R^2 = 0.8108$ for the training set, and $y = 1.1507x - 1.4149$, $R^2 = 0.814$ for the test set; (b) W1 only, $y = 0.6703x + 2.7151$, $R^2 = 0.6703$ for the training set, and $y = 1.484x - 4.1632$, $R^2 = 0.7287$ for the test set; and (c) 2D descriptors only, $y = 1x + 1 \times 10^{-5}$, $R^2 = 0.7485$ for the training set, and $y = -0.6951x + 13.551$, $R^2 = 0.357$ for the test set.

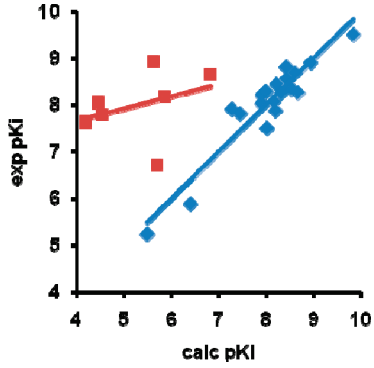

Figure 4. Correlations of experimental and calculated pK_i's in QSAR models for trypsin generated on the training set (blue) and examined on the test set (red). The training set contains 19 compounds: 1b, 4, 51−53, and 55−68 from Table 5 in ref 23a. The test set consists of 7 compounds: 2b−d, 2f, and 2h−j from Table 1 in ref 23b. W1 and 2D descriptors, $y = 1x + 7 \times 10^{-6}$, $R^2 = 0.8869$ for the training set, and $y = 0.2634x + 6.6056$, $R^2 = 0.1181$ for the test set.

are equally well adopted for thrombin. Indeed compounds with the same CS, compound number 57 from the first and 2a from the second series, have a very similar K_i in the nanomolar range. The role of RS is to optimally align the CS fragment in the enzyme active site, assuring the best positioning of its reactivity center for interaction with the attacking enzyme nucleophile. The RS of the second series used as a test set is far from being optimized to trypsin: K_i values of compounds with the same CS are markedly different −8 μM for 2a in the test set vs 1.2 nM for 57 in the training set. Thus, it is not surprising that the QSAR model calibrated on the optimal alignment of CS fragments with RS selective to trypsin in the training set failed to predict for a nonspecific RS in the test set, where positioning of CS in the trypsin active site is nonoptimal and different from that in the training set.

In support of this analysis, we split the series of inhibitors with optimal RS, formerly used as a training set,[23a] into two subsets: 14 compounds for a training set and 5 compounds as a test set. Now the prediction by the QSAR model containing W1 and four 2D noncovalent descriptors is quite good: $R^2 = 0.835$ (Figure 5a and Table S2 in the Supporting Information), compatible with the thrombin case. Sequential exclusion of variables demonstrated that W1 plays a key role also in the QSAR predictability in trypsin (Figure 5b, Table S2 in the Supporting Information), and the prediction based on noncovalent descriptors only is absolutely irrelevant (Figure 5c and Table S2 in the Supporting Information). We conclude from the above comparison of thrombin and trypsin that a QSAR model for prediction of CS binding trend must be calibrated on a training set and applied to a set with a RS

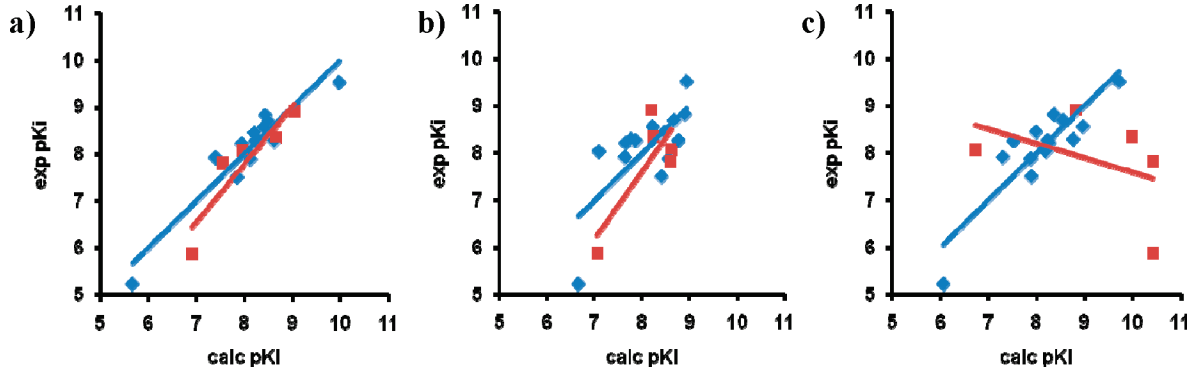

Figure 5. Correlations of experimental and calculated pK_i's in QSAR models for trypsin generated on the training set (blue) and examined on the test set (red). The training set contains 14 compounds: 1b, 4, 52, 53, 55, 58−60, and 63−68, and the test set consists of 5 compounds: 51, 56, 57, 61, and 62 from Table 5 in ref 23a. (a) W1 and 2D descriptors, $y = 1x + 2 \times 10^{-5}$, $R^2 = 0.8903$ for the training set, and $y = 1.2304x - 2.0647$, $R^2 = 0.8349$ for the test set; (b) W1 only, $y = 1x + 8 \times 10^{-6}$, $R^2 = 0.5197$ for the training set, and $y = 1.4391x - 3.907$, $R^2 = 0.633$ for the test set; (c) 2D descriptors only, $y = 1x + 2 \times 10^{-5}$, $R^2 = 0.7781$ for the training set, and $y = -0.3062x + 10.651$, $R^2 = 0.1726$ for the test set.

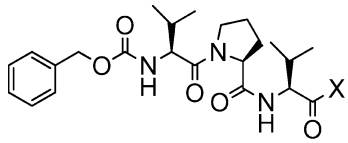

Figure 6. Peptidyl α-ketoheterocyclic inhibitors of HNE. X is a varied heterocycle. The training set contains 26 compounds: a, f−j, l, and p−s from Table 1, and 7 and 21−25 from Table 3 in ref 24a and 6b−d, 6f−j, 6l from Table 1 in ref 24b. The test set consists of 7 compounds: b, e, k, and o from Table 1 and 26 from Table 3 in ref 24a and 6e and 6k from Table 1 in ref 24b.

fragment selective to the target enzyme. In other words, the prediction of optimal CS by the QSAR model is relevant only if the inhibitor contains a RS specific for the target enzyme.

HNE. A series of isoselective compounds collected from two literature sources (Figure 6)[24] was used for QSAR analysis of RCA inhibitors of the serine protease human neutrophil elastase (HNE). The series was split into a training set of 26 compounds and a test set of 7 compounds. The QSAR model optimized by GFA method identified W1 and four 2D descriptors as the best variant, with prediction on the test set that is characterized by $R^2 = 0.933$.

Sequential exclusion of variables (Figure 7 and Table S3 in the Supporting Information) demonstrates that W1 plays a key role in the QSAR predictability in HNE. Prediction with noncovalent descriptors only is poor.

FAAH. The QSAR model for the serine hydrolase FAAH was generated by GFA algorithm on quadratic model containing five independent variables: W1 and four functions characterizing noncovalent interactions of CS with the enzyme active site. The model was built on a training set of 20 compounds and validated for prediction with $R^2 = 0.811$ on a test set of 6 compounds. Compounds for both sets were chosen from a series of isoselective RCA inhibitors, collected from two literature sources (Figure 8).[25] Sequential exclusion of variables demonstrates that there is a synergistic effect of W1 and the noncovalent variables, having good prediction only when combined (Figure 9a and Table S4 in the Supporting Information). W1 alone is inadequate for the correlation on the training and test sets (Figure 9b and Table S4 in the Supporting Information). We can speculate that the carbonyl group, the reactivity center of α-ketoheterocyclic CS fragments, is not optimally aligned for the chemical

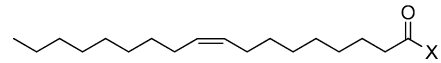

Figure 8. Oleoyl α-ketoheterocyclic inhibitors of FAAH. X is a varied heterocycle. The training set contains 20 compounds: 3, 4, 6, 7, 12−14, 17, 18, 20, and 23 from Table 1 and 29, 30, 32 from Table 3 in ref 25a and 64−66, 70−72 from Table 7 in ref 25b. The test set consists of 6 compounds: 5, 10, 11, 19 from Table 1 and 31 from Table 3 in ref 25a and 67 from Table 7 in ref 25b.

interaction with the attacking Ser nucleophile in the active site of FAAH by the oleoyl fragment (Figure 8). Indeed it was noted that the inhibitor potency progressively increased as the fatty acid chain length was shortened from C_{18} to C_{12}.[25a] This statement is supported by a low value of $R^2 = 0.2622$ calculated for the test set in the QSAR model built on 2D descriptors only (Figure 9c and Table S4 in the Supporting Information). Further analysis of the reasons for different selectivity of RS fragments of FAAH RCA inhibitors needs comprehensive 3D molecular modeling studies on the whole protein, applying molecular dynamics and other relevant approaches, which is far beyond the scope of the present study.

For all of the serine hydrolases studied, the GFA algorithm selected W1 as the QSAR descriptor accounting for the enzyme−inhibitor covalent interaction contribution to the RCA inhibitors binding trend. W1 descriptor is sensitive to the specificity of the RC fragment controlling the quality of the alignment of the CS in the enzyme active site. Accounting for both covalent and noncovalent interactions of CS in a combined QSAR model built on W1 and 2D descriptors provides the best correlation between experimental and predicted binding trend of RCA inhibitors.

Cathepsin K. We recently demonstrated by molecular modeling of papain inhibition[29] and QSAR analysis of cathepsin K[13] that the weak Cys nucleophile in cysteine proteases forms a neutral TC(OH) covalent tetrahedral complex with carbonyl groups of RCA inhibitors. This is further supported by our previous theoretical studies of thiolate reactivity toward different electrophiles[30] and by the conclusions derived from experimental kinetic studies of papain.[31] We also demonstrated that with a weak Cys nucleophile in cathepsin K, the binding trend of CS of RCA inhibitors is dominated by the change in pK_a of the oxygen atom in TC(OH), which is accounted for by the covalent descriptor W2.[13] The QSAR models presented in this work

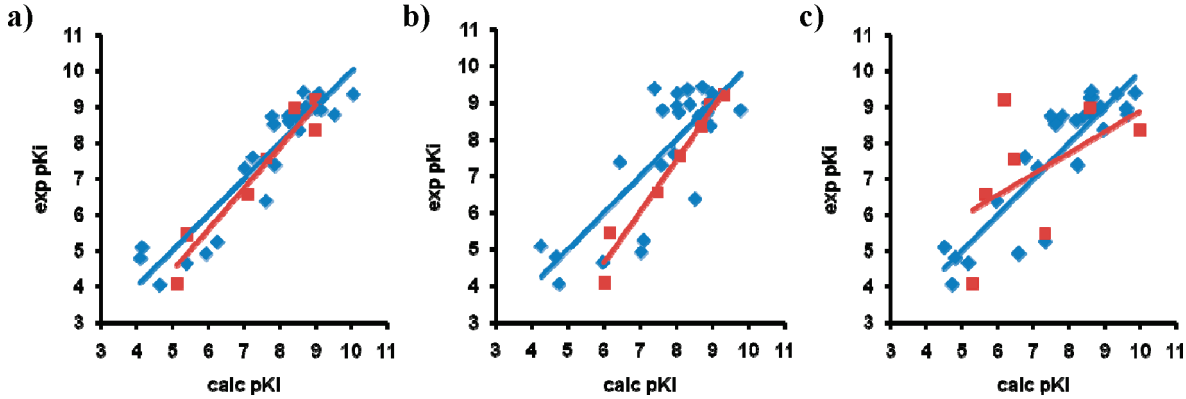

Figure 7. Correlations of experimental and calculated pK_i's in QSAR models for HNE generated on 26 compounds in the training set (blue) and examined on 7 compounds in the test set (red). (a) W1 and 2D descriptors, $y = 1x + 9 \times 10^{-6}$, $R^2 = 0.8723$ for the training set, and $y = 1.1475x - 1.2994$, $R^2 = 0.9331$ for the test set; (b) W1 only, $y = 1x + 3 \times 10^{-6}$, $R^2 = 0.6474$ for the training set, and $y = 1.4136x - 3.8685$, $R^2 = 0.9644$ for the test set; and (c) 2D descriptors only, $y = 1x - 2 \times 10^{-8}$, $R^2 = 0.7909$ for the training set, and $y = 0.5839x + 3.0472$, $R^2 = 0.2692$ for the test set.

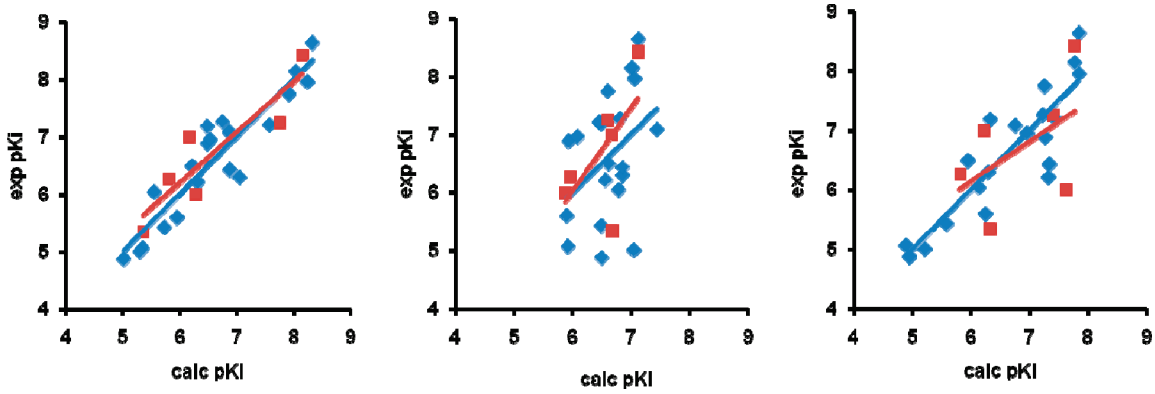

Figure 9. Correlations of experimental and calculated pK_i's in QSAR models for FAAH generated on 20 compounds in the training set (blue) and examined on 6 compounds in the test set (red). (a) W1 and 2D descriptors, $y = 1x - 6 \times 10^{-6}$, $R^2 = 0.8649$ for the training set, and $y = 0.8791x + 0.9219$, $R^2 = 0.8114$ for the test set; (b) W1 only, $y = 1x + 1 \times 10^{-5}$, $R^2 = 0.157$ for the training set, and $y = 1.4293x - 2.5508$, $R^2 = 0.3878$ for the test set; and (c) 2D descriptors only, $y = 1x - 1 \times 10^{-6}$, $R^2 = 0.7765$ for the training set, and $y = 0.6649x + 2.1549$, $R^2 = 0.2622$ for the test set.

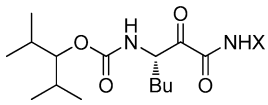

Figure 10. Reversible covalent ketoamide inhibitors of cathepsin K with varied substituent X. Training set compounds: 13, 14, 18−20, 22, 24−32, 34−40, and 42 and test set: 15, 16, 21, 23, 33, 41, and 43.[26]

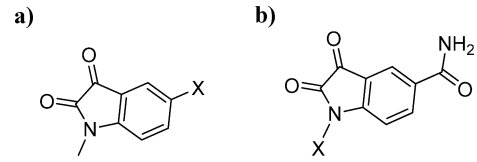

Figure 12. Isatin derivatives with varied substituent X at: (a) C5 and (b) N1 positions, as RCA inhibitors of HRV 3C cysteine protease.[27]

are generated for reversible covalent ketoamide inhibitors of cathepsin K (Figure 10).[26] The training set contains 23 compounds, and the test set contains 7 compounds.

QSAR models generated by GFA algorithm may contain both binary interactions and simple quadratic terms in addition to linear terms.[16,21] The best QSAR model was a quadratic model containing three variables, one of which is a product of two of the initial descriptors (Table S5 in the Supporting Information). This model is based on the combined set of descriptors W2 and 2D with $R^2 = 0.9239$ calculated for the prediction on the test set (Figure 11a). Effective design of CS requires a wide reactivity interval of the training set. In ref 26 the varied X fragments are attached to the reactivity center of inhibitors by a spacer. Therefore, the varied X can only moderately modify the electron density on CS and therefore its reactivity. This is reflected by a low $R^2 = 0.6429$ for the prediction on the test set for a QSAR model optimized on W2 alone (Figure 11b and Table S5 in the Supporting Information). On the other hand, the flexible X fragments participate in various polar and hydrophobic noncovalent interactions in the enzyme active site. Indeed, the value of $R^2 = 0.7529$ for the prediction on a test set by a QSAR model built on 2D descriptors reflects the considerable role of these noncovalent interactions of CS (Figure 11c and Table S5 in the Supporting Information).

HRV 3C. Two sets of isatin derivatives with varied substituent X at the C5 and N1 positions were used in the QSAR analysis of RCA inhibitors of the human rhinovirus (HRV) 3C cysteine protease (Figure 12a and b, respectively).[27] The size of both sets of compounds is too small (8 and 14 compounds, respectively) for separation into training and test sets, so we only analyzed the correlation pK_i^{exp} vs pK_i^{calc} (Figures 13 and 14).

The catalytic Cys147 of HRV 3C forms a covalent bond with the electrophilic C3 position of isatin, yielding a tetrahedral complex TC(OH), in analogy to the cathepsin K−inhibitor complex.[27] The varied substituent X at C5

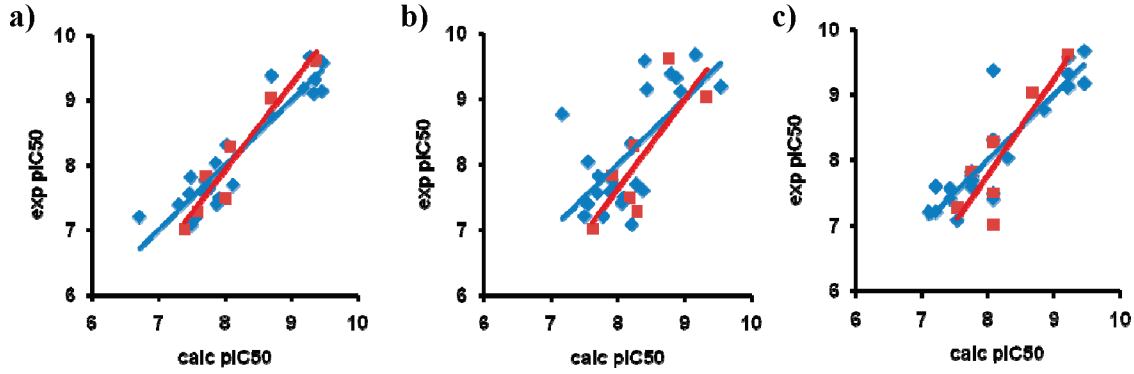

Figure 11. Correlations of experimental and calculated pIC50 in QSAR models for cathepsin K, generated on 20 compounds in the training set (blue) and examined on 7 compounds in the test set (red). (a) W2 and 2D descriptors, $y = 1x + 1 \times 10^{-5}$, $R^2 = 0.8693$ for the training set, and $y = 1.3236x - 2.6619$, $R^2 = 0.9239$ for the test set; (b) W2 only, $y = 1x + 1 \times 10^{-5}$, $R^2 = 0.4461$ for the training set, and $y = 1.3729x - 3.3556$, $R^2 = 0.6429$ for the test set; and (c) 2D descriptors only, $y = 1x + 8 \times 10^{-6}$, $R^2 = 0.7998$ for the training set, and $y = 1.4665x - 3.9519$, $R^2 = 0.7529$ for the test set.

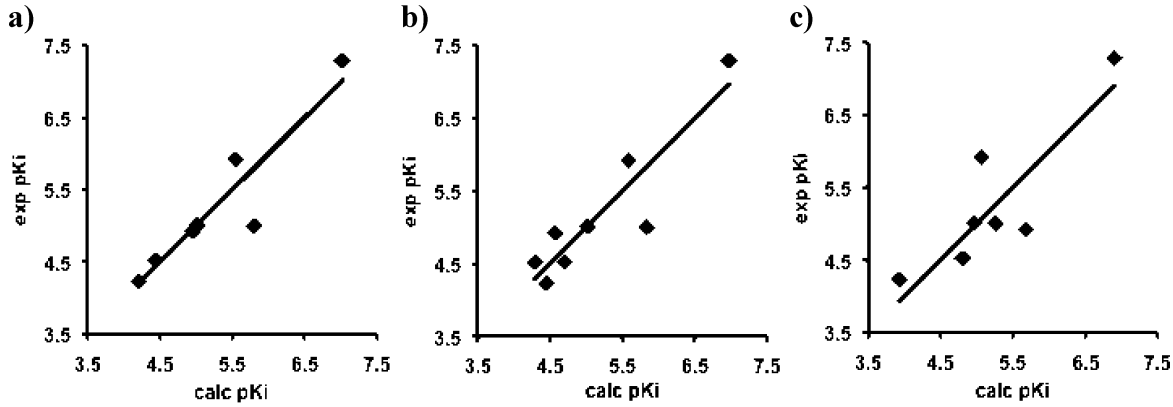

Figure 13. Correlations of experimental and calculated pK_i's in QSAR models for HRV 3C, generated on 8 compounds, 2−6, 8, 9, and 12 (Figure 12a).[27] (a) W2 and 2D descriptors, $y = 1x - 6 \times 10^{-6}$, $R^2 = 0.8739$; (b) W2 only, $y = 1x - 6 \times 10^{-6}$, $R^2 = 0.8311$; and (c) 2D descriptors only, $y = 1x + 1 \times 10^{-5}$, $R^2 = 0.7427$.

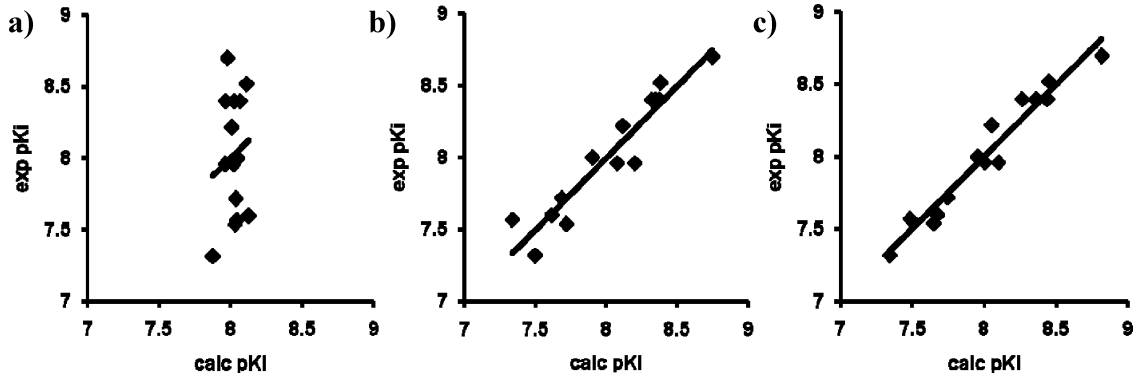

Figure 14. Correlations of experimental and calculated pK_i's in QSAR models for HRV 3C generated on 14 compounds, 14−27.[27] (a) W2 descriptor, $y = 1x + 4 \times 10^{-4}$, $R^2 = 0.0223$; b) 2D descriptors, $y = 1x - 6 \times 10^{-6}$, $R^2 = 0.8988$; and (c) W2 and 2D descriptors, $y = 1x - 4 \times 10^{-6}$, $R^2 = 0.9508$.

(Figure 12a) modifies the electron density on the isatin ring and thus influences the electrophilicity of the C3 atom and the proton affinity of the corresponding carbonyl oxygen atom. QSAR models for HRV 3C were optimized by GFA algorithm for three variants of independent variables: W2 and 2D descriptors, W2 only, and 2D (Figure 13 and Table S6 in the Supporting Information). Two of the QSAR models, built on W2 and 2D descriptors, have close values of $R^2 = 0.8311$ and 0.7427, respectively. This means that both factors—pK_a of TC(OH) accounted for by W2 and the CS enzyme−inhibitor noncovalent interactions accounted for by 2D descriptors—are approximately equally significant for the binding trend of this set of inhibitors. As expected, the combined W2 and 2D model improves the correlation, $R^2 = 0.8739$, where W2 plays a considerable role.

It is worth noting that W2 has a positive coefficient in the QSAR models calculated for the cysteine proteases cathepsin K and HRV 3C, in contrast to the negative coefficients of W1 in all of the considered QSAR models for serine proteases. As rationalized previously, the phenomenon originates in the opposite trends of electrophilicity of the carbon atom and the proton affinity of the oxygen atom of the carbonyl group caused by the variation of the X substituent in the CS of RCA inhibitors.[13]

The varied X substituents at the isatin N1 position in the second series of inhibitors (Figure 12b) are separated from the isatin π-conjugated electronic system by a methylene linker.[27] Therefore, the changes in electron density in the isatin ring should be minor. Indeed, the QSAR model based on W2 descriptor only (Figure 14a and Table S7 in the Supporting Information) shows a very narrow interval for the change of calculated pK_i. This demonstrates the relevance of the covalent W2 descriptor, which correctly reflects the negligible change in electron density on the reactivity center. The calculated very small $R^2 = 0.0223$ indicates a negligible role for the electronic factor in the binding trend of this set of inhibitors. It is interesting to note that the linear regression coefficient at W2 is now negative, in contrast to the positive coefficients in the model of cathepsin K and the first inhibitor series of HRV 3C. This stems from the very small electronic effects in the second series of inhibitors of HRV 3C. The W2 descriptor reflects two electronic factors (see also eqs 1and 2):[13] the strengths of the enzyme−inhibitor covalent bond in TC(OH) and the pK_a of the latter. The negative coefficient at W2 means that the factor of small variation of the enzyme−inhibitor covalent bond strength outweighs the negligible change in pK_a in this specific case. The X substituents in this series of inhibitors are flexible, quite large extended molecular fragments,[27] so it is not surprising why 2D descriptors, which account for the noncovalent interactions with the enzyme active site, are the trend dominating with $R^2 = 0.8988$ (Figure 14b and Table S7 in the Supporting Information). The QSAR model built on the combined set of independent variables W2 and 2D (Figure 14c and Table S7 in the Supporting Information) demonstrates a synergistic improvement, $R^2 = 0.9508$, in the correlation of experimental pK_i vs calculated pK_i.

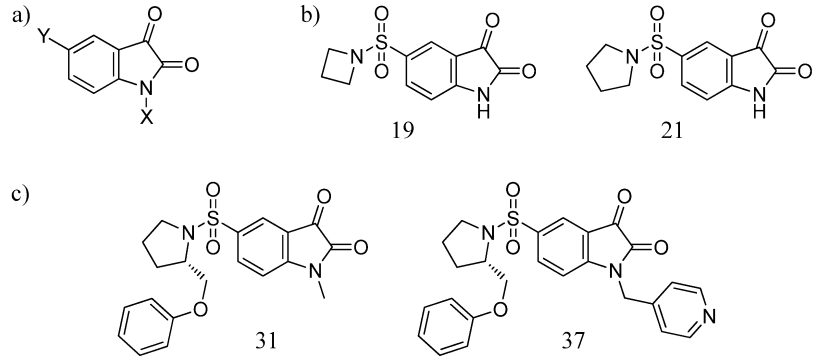

Figure 15. Isatin sulfonamide RCA inhibitors of caspases 3 with varied substituents: (a) X in position N1 and Y in position C5; (b) Examples of the variation of the Y fragment, compounds 19 and 21; and (c) Examples of the variation of the X fragment, compounds 31 and 37.[28]

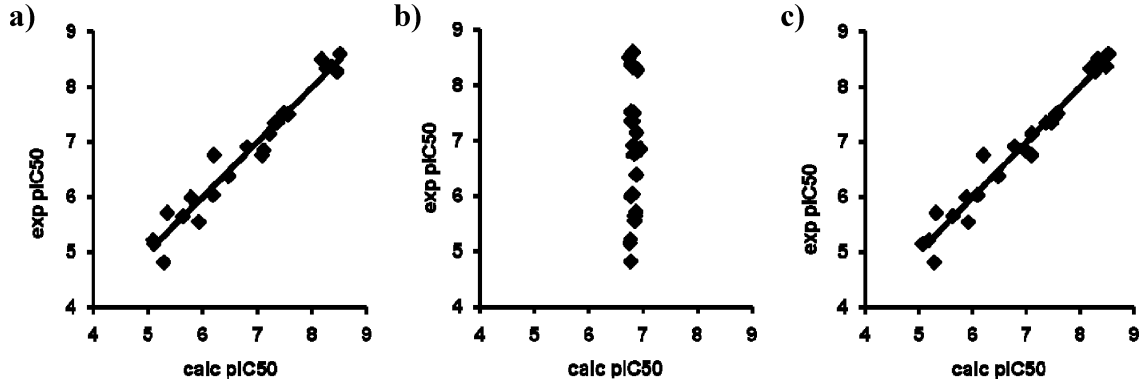

Figure 16. Correlations of experimental and calculated pIC50s in QSAR models for caspase 3 generated on 23 compounds: 1, 5−7, 16, 18−26, 28−35, and 37.[28] a) 2D descriptors, $y = 1x - 5 \times 10^{-6}$, $R^2 = 0.9572$; (b) W2 descriptor, $y = 1x + 2 \times 10^{-4}$, $R^2 = 0.0019$; and (c) W2 and 2D descriptors, $y = 1x + 4 \times 10^{-6}$, $R^2 = 0.963$.

Caspase 3. We could not find in the literature any statistically representative set of isoselective RCA inhibitors of the cysteine protease caspase 3. Nevertheless we found two sets of isatin sulfonamide inhibitors, substituted at the N1 and C5 positions (Figure 15).[28] QSAR analysis of the full set of 23 compounds from both sets is presented in Figure 16 and Table S8 in the Supporting Information.

The best QSAR model ($R^2 = 0.9572$) identified by the GFA algorithm contains five 2D descriptors only but not W2 (Figure 16a and Table S8 in the Supporting Information). This is not surprising since the varied X and Y substituents are not conjugated to the reaction center of the isatin ring. Moreover, they are flexible and with different steric properties,[28] and therefore the binding trend is dominated by noncovalent enzyme−inhibitor interactions. W2 plays a negligible role in binding trend of the series of caspase 3 inhibitors (Figure 16b and Table S8 in the Supporting Information), in analogy to the case of HRV 3C (Figure 14a). We have also constructed the combined QSAR model built by multiple linear regression on W2 and the 2D descriptors (Figure 16c and Table S8 in the Supporting Information). The model demonstrates that the W2 descriptor does not improve the correlation. Finally we note that W2 has negative correlation coefficients (Table S8 in the Supporting Information) by the same reasons discussed for the second set of HRV 3C inhibitors.

W2 was identified by GFA as a key factor for QSAR analysis of the binding trend of series of RCA inhibitors of cysteine proteases if the varied substituents significantly influence the electron density and therefore the reactivity of the carbonyl reactivity center. In such cases the pK_a of TC(OH) considerably contributes to the binding trend. W2 plays a minor role if the varied substituent in CS does not modulate the electron density on the reactivity center. In the latter case the noncovalent interactions of CS with the enzyme active site dominate the binding trend, and a good QSAR model can be built on 2D descriptors only.

CONCLUSIONS

We have introduced an enzyme mechanism-based method (EMBM) for rational design of chemical sites of reaction coordinate analog inhibitors. The energy of valence reorganization of chemical sites (CS) upon the enzyme−inhibitor covalent complex formation is accounted for by new covalent descriptors W1 and W2. We considered CS fragments with a carbonyl reactivity center, similar to native substrates of proteases. The W1 and W2 descriptors are calculated quantum mechanically on small molecular clusters simulating the reaction core of the formed covalent tetrahedral complex, either anionic TC(O−) or neutral TC(OH). The modeling allows the generation of various CS and corresponding TC(O−) and TC(OH) as universal building blocks of inhibitors and their covalent complexes with serine or cysteine hydrolases.

We have demonstrated that two factors influence the binding affinity of CS fragments. The first is the energy of reorganization of covalent bonds of the CS fragment during the chemical transformation in the enzyme active site, accounted by the W1 and W2 covalent descriptors. Every CS fragment also participates in differenet noncovalent interactions in the enzyme active site, which may result in

conformational changes of the CS fragment in its bound state in comparison with the free ligand. These noncovalent interactions of CS as well as the resulting conformational changes during the alignment in the enzyme active site are individual for every enzyme. This was demonstrated in the comparative examples of thrombin and trypsin. The contribution of noncovalent interactions of CS fragments to the binding affinity of inhibitors and the energy of their conformational changes are accounted for in this work by 2D descriptors. The relative role of the covalent and noncovalent interactions depends on different factors of the CS structure: the influence of the varied part of CS on the electronic structure of the reaction core of the inhibitor, the size and conformational flexibility, participation in H-bonds, hydrophobicity, etc. Different variants of such interplay between covalent and noncovalent interactions of CS fragments, reflected in their W1 and W2 covalent and 2D noncovalent descriptors, were analyzed. Neither the 2D nor the W1 and W2 descriptors requires knowledge of the 3D structure of the target enzyme, so EMBM may be used for ligand-based design, a variant of which is considered in the present study.

We have generated a chemical site of inhibitors (CSI) databank, with CS fragments and corresponding pairs of W1 and W2 descriptors precalculated for both $CH_3O(-)$ and $CH_3S(-)$ nucleophiles at a wide range of different dielectric constants. A varied substituent in CS has an opposite influence on the electrophilicity of the carbonyl carbon atom and the pK_a of the former carbonyl oxygen atom in the TC(OH). The former property dominates the binding trend of RCA inhibitors with serine hydrolases and the latter with cysteine hydrolases. Therefore, W1 is an optimal covalent descriptor for serine and W2 for cysteine hydrolases. The EMBM was validated on a few serine proteases, thrombin, trypsin, and human neutrophil elastase (HNE); a serine hydrolase fatty acid amide hydrolase (FAAH); and cysteine proteases: cathepsin K, HRV 3C, and caspase 3. The W1 and W2 descriptors were used as precalculated values from our ICS databank. CS of reaction-coordinate analogue (RCA) inhibitors can be designed separately on the modeling set of isoselective inhibitors (with a constant RS) and then applied to another set of inhibitors with a different RS specific for the target enzyme. EMBM can be used in combination with other ligand- or structure-based computer-assisted drug design (CADD) tools for optimal design of inhibitors.

EMBM may help in overcoming mutational drug resistance, as it concentrates on the nonvariable, not subject to mutations, catalytic machinery of the target enzyme. EMBM will provide an easy-to-use, accurate, and cost-effective tool for drug design.

ACKNOWLEDGMENT

This research was supported by NIH grant GM081329.

Supporting Information Available: Tables S1−S8 contain description of the QSAR models of the various enzymes. Table S9 contains description of the noncovalent 2D QSAR descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(2) Keserü, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discovery* **2009**, *8*, 203–212.

(3) (a) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676. (b) Yang, G. F.; Huang, X. Development of quantitative structure-activity relationships and its application in rational drug design. *Curr. Pharm. Des.* **2006**, *12*, 4601–4611.

(4) Houk, K. N.; Leach, A. G.; Kim, S. P.; Zhang, X. Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. *Angew. Chem., Int. Ed.* **2003**, *42*, 4872–4897.

(5) Wolfenden, R.; Snider, M. J. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **2001**, *34*, 938–945.

(6) (a) Wolfenden, R. Transition state analogues for enzyme catalysis. *Nature* **1969**, *223*, 704–705. (b) Schramm, V. L. Enzymatic transition states and transition state analogues. *Curr. Opin. Struct. Biol.* **2005**, *15*, 604–613. (c) Zhang, X.; Houk, K. N. Why enzymes are proficient catalysts: Beyond the Pauling paradigm. *Acc. Chem. Res.* **2005**, *38*, 379–385.

(7) Robertson, J. G. Mechanistic basis of enzyme-targeted drugs. *Biochemistry* **2005**, *44*, 5561–5571.

(8) (a) Warshel, A. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *11*, 5250–5254. (b) Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. Electrostatic basis for enzyme catalysis. *Chem. Rev.* **2006**, *106*, 3210–3235.

(9) Maryanoff, B. E.; Costanzo, M. J. Inhibitors of proteases and amide hydrolases that employ an alpha-ketoheterocycle as a key enabling functionality. *Bioorg. Med. Chem.* **2008**, *16*, 1562–1595.

(10) Christianson, D. W.; Lipscomb, W. N. Comparison of carboxypeptidase-A and thermolysin - inhibition by phosphonamidates. *J. Am. Chem. Soc.* **1988**, *110*, 5560–5565.

(11) Shokhen, M.; Arad, D. Anionic tetrahedral complexes as serine protease inhibitors. *J. Mol. Model.* **1996**, *2*, 390–398.

(12) (a) Ozeri, R.; Khazanov, N.; Perlman, N.; Shokhen, M.; Albeck, A. Enzyme isloselective inhibitors: A tool for binding-trend analysis. *ChemMedChem* **2006**, *1*, 631–638. (b) Shokhen, M.; Khazanov, N.; Albeck, A. Enzyme isoselective inhibitors: Application to drug design. *ChemMedChem* **2006**, *1*, 639–643.

(13) Shokhen, M.; Traube, T.; Vijayakumar, S.; Hirsch, M.; Uritsky, N.; Albeck, A. Differentiating serine and cysteine proteases mechanism by new covalent QSAR descriptors. 2010, submitted.

(14) (a) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage 0.1. Validation and estimation of coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285–294. (b) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage 0.2. Application of conformational models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295–304. (c) Smellie, A.; Teig, S. L.; Towbin, P. Poling - promoting conformational variation. *J. Comput. Chem.* **1995**, *16*, 171–187.

(15) Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. Derivation of class-ii force-fields 0.1. Methodology and quantum force-field for the alkyl functional-group and alkane molecules. *J. Comput. Chem.* **1994**, *15*, 162–182.

(16) *Discovery Studio Modeling Environment*, release 2.5; Accelrys Software Inc.: San Diego, CA, 2009.

(17) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(18) Stewart, J. J. P. *MOPAC 2009*; Stewart Computational Chemistry, Colorado Springs, CO, 2008; http://OpenMOPAC.net.

(19) Shokhen, M.; Khazanov, N.; Albeck, A. Screening of the active site from water by the incoming ligand triggers catalysis and inhibition in serine proteases. *Proteins* **2008**, *70*, 1578–1587.

(20) Klamt, A.; Schüürmann, G. COSMO - a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799–805.

(21) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(22) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

(23) (a) Costanzo, M. J.; Almond Jr, H. R.; Hecker, L. R.; Schott, M. R.; Yabut, S. C.; Zhang, H. C.; Andrade-Gordon, P.; Corcoran, T. W.; Giardino, E. C.; Kauffman, J. A.; Lewis, J. M.; de Garavilla, L.; Haertlein, B. J.; Maryanoff, B. E. In-depth study of tripeptide-based alpha-ketoheterocycles as inhibitors of thrombin. Effective utilization of the S-1' subsite and its implications to structure-based drug design. *J. Med. Chem.* **2005**, *48*, 1984–2008. (b) Bachand, B.; Tarazi, M.; St-Denis, Y.; Edmunds, J. J.; Winocour, P. D.; Leblonda, L.;

Siddiqui, M. A. Potent and selective bicyclic lactam inhibitors of thrombin. Part 4: Transition state inhibitors. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 287–290.

(24) (a) Edwardsp, P. D.; Wolanin, D. J.; Andisik, D. W.; Davis, M. W. Peptidyl alpha-ketoheterocyclic inhibitors of human neutrophil elastase. 2. Effect of varying the heterocyclic ring on in-vitro potency. *J. Med. Chem.* **1995**, *38*, 76–85. (b) Edwards, P. D.; Zottola, M. A.; Davis, M.; Williams, J.; Tuthillt, P. A. Peptidyl alpha-ketoheterocyclic inhibitors of human neutrophil elastase. 3. In-vitro and in-vivo potency of a series of peptidyl alpha-ketobenzoxazoles. *J. Med. Chem.* **1995**, *38*, 3972–3982.

(25) (a) Boger, D. L.; Sato, H.; Lerner, A. E.; Hedrick, M. P.; Fecik, R. A.; Miyauchi, H.; Wilkie, G. D.; Austin, B. J.; Patricelli, M. P.; Cravatt, B. F. Exceptionally potent inhibitors of fatty acid amide hydrolase: The enzyme responsible for degradation of endogenous oleamide and anandamide. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5044–5049. (b) Seierstad, M.; Breitenbucher, J. G. Discovery and Development of Fatty Acid Amide Hydrolase (FAAH) Inhibitors. *J. Med. Chem.* **2008**, *51*, 7327–7343.

(26) Tavares, F. X.; Boncek, V.; Deaton, D. N.; Hassell, A. M.; Long, S. T.; Miller, A. B.; Payne, A. A.; Miller, L. R.; Shewchuk, L. M.; Wells-Knecht, K.; Willard Jr., D. H.; Wright, L. L.; Zhou, H.-Q. Design of potent, selective, and orally bioavailable inhibitors of cysteine protease cathepsin K. *J. Med. Chem.* **2004**, *47*, 588–599.

(27) Webber, S. E.; Tikhe, J.; Worland, S. T.; Fuhrman, S. A.; Hendrickson, T. F.; Matthews, D. A.; Love, R. A.; Patick, A. K.; Meador, J. W.; Ferre, R. A.; Brown, E. L. D.; DeLisle, M.; Ford, C. E.; Binford, S. L. Design, synthesis, and evaluation of nonpeptidic inhibitors of human rhinovirus 3C protease. *J. Med. Chem.* **1996**, *39*, 5072–5082.

(28) Lee, D.; Long, S. A.; Murray, J. H.; Adams, J. L.; Nuttall, M. E.; Nadeau, D. P.; Kikly, K.; Winkler, J. D.; Sung, C.-M.; Ryan, M. D.; Levy, M. A.; Keller, P. M.; DeWolf, Jr, W. E. Potent and selective nonpeptide inhibitors of caspases 3 and 7. *J. Med. Chem.* **2001**, *44*, 2015–2026.

(29) Shokhen, M.; Khazanov, N.; Albeck, A. The mechanism of papain inhibition by peptidyl aldehydes. *Proteins* **2010**, in press.

(30) Shokhen, M.; Arad, D. The source for the difference between sulfhydryl and hydroxyl anions in their nucleophilic addition reaction to a carbonyl group: A DFT approach. *J. Mol. Model.* **1996**, *2*, 399–409.

(31) Frankfater, A.; Kuppy, T. Mechanism of association of n-acetyl-l-phenylalanylglycinal to papain. *Biochemistry* **1981**, *20*, 5517–5524.

CI100330Y