

Improving Quantitative Structure–Activity Relationships through Multiobjective Optimization

Orazio Nicolotti,* Ilenia Giangreco, Teresa Fabiola Miscioscia, and Angelo Carotti

Dipartimento Farmaco-Chimico, University of Bari, via Orabona 4, I-70125 Bari, Italy

Received July 8, 2009

A multiobjective optimization algorithm was proposed for the automated integration of structure- and ligand-based molecular design. Driven by a genetic algorithm, the herein proposed approach enabled the detection of a number of trade-off QSAR models accounting simultaneously for two independent objectives. The first was biased toward best regressions among docking scores and biological affinities; the second minimized the atom displacements from a properly established crystal-based binding topology. Based on the concept of dominance, 3D QSAR equivalent models profiled the Pareto frontier and were, thus, designated as nondominated solutions of the search space. K-means clustering was, then, operated to select a representative subset of the available trade-off models. These were effectively subjected to GRID/GOLPE analyses for quantitatively featuring molecular determinants of ligand binding affinity. More specifically, it was demonstrated that a) diverse binding conformations occurred on the basis of the ligand ability to profitably contact different part of protein binding site; b) enzyme selectivity was better approached and interpreted by combining diverse equivalent models; and c) trade-off models were successful and even better than docking virtual screening, in retrieving at high sensitivity active hits from a large pool of chemically similar decoys. The approach was tested on a large series, very well-known to QSAR practitioners, of 3-amidinophenylalanine inhibitors of thrombin and trypsin, two serine proteases having rather different biological actions despite a high sequence similarity.

INTRODUCTION

Molecular discovery requires the simultaneous optimization of diverse and often conflicting properties (i.e., objectives). A few examples are molecular diversity and drug-likeness in designing combinatorial libraries;¹ molecular alignments and pharmacophore definition in substructure mining; accuracy and structural complexity in Quantitative Structure–Activity Relationships (QSAR) modeling; docking score and rmsd in molecular design; and optimization of ADME properties in drug discovery.^{2,3} Despite this real-life evidence, standard methods ignore this issue preferring stepwise optimization procedures. Known as single-objective optimization (SOO), these approaches implicitly assume that even for highly complex problems only one single optimal solution exists. Unlike SOO, multiobjective optimization (MOO) techniques are aimed at finding trade-offs among the various objectives and, thus, at identifying not one but a pool of equivalent models. Despite the enormous potential,⁴ MOO techniques have been only recently applied to QSAR and, actually, few articles exist.^{5,6} Some of us and coll. contributed to report the first application of MOO to QSAR.⁸ It employed a variant of an evolutionary algorithm called multiobjective genetic programming using Pareto ranking to optimise a number of conflicting objectives including model accuracy, number of terms, internal complexity, and interpretability of the descriptors occurring in 2D QSAR model.⁹ Stockfish and coll. proposed a nonevolutionary multiobjective technique to construct models to analyze selectivity relationships between cyclooxygenases 1 and 2 inhibitors.¹⁰

More recently, Cruz-Montegudo and coll. proposed a MOO method based on the Derringer's desirability function that allows running global QSAR studies jointly considering multiple properties of interest in drug design process.¹¹ An excellent review recently published by Nicolaou and coll. surveys the last advancement in drug discovery through MOO.¹²

The present investigation represents the MOO evolution of our recent paper aimed at selecting appropriate docking poses for building 3D QSAR.¹³ Strictly speaking, the herein proposed analysis is a dual-objective optimization as trade-offs were effectively searched on the basis of two different terms. Unlike previous strategies basically considering top-scored docking poses supposed to represent the true ligand binding mode¹⁴ or, alternatively, choosing via cluster analysis poses having low root-mean-square deviations (rmsd), an automatic method was conceived which makes use of a genetic algorithm optimising a weighted sum fitness function composed by two terms. The first computed the correlation between docking scores and experimental biological activity, and the second took into account the averaged rmsd with respect to a defined cocrystallized inhibitor. The method was successfully applied to the analysis of a well-known series of 88 benzamidine inhibitors (72 in the training and 16 in the test set, respectively) of thrombin and trypsin intensely examined also by other authors with different approaches.^{15,16} As published, the method offered the advantage of obtaining independent QSAR models for diverse biological targets and, thus, represented a useful investigative tool for analyzing molecular recognition and selectivity.^{13,17} The occurrence of nonequivalent interactions of diversely oriented inhibitors

* Corresponding author e-mail: nicolotti@farmchim.uniba.it.

into the active sites of two distinct proteins was in fact explicitly considered. In addition, the procedure permitted to significantly reduce the propagation of the experimental uncertainty of each independent biological measure as no difference of pK_i relative to diverse biological targets were ever taken into account.

However, our previous approach was based on a SOO method which normally strives to discover a single optimal solution for complex optimization problems. It was supposed that the weighted sum fitness function considered both terms linearly combined and equally important to the problem. Tacitly, it was assumed that identical coefficients were associated with both terms, or, alternatively, a calibration procedure was required for changing them to more desirable values for a given user. Although this option, the final outcome of our previous approach, likewise all the SOO methods, still remained one single solution. In the present work, the two terms initially included in the weighted sum fitness function were transformed into objectives themselves and both subjected to the MOO. This change permitted to have a consistent gain in optimality through compromises and trade-offs between the two terms (here referred to as objectives) and, thus, to result in a family of equivalent QSAR models. The user is, hence, given the option to compare and choose models which better explain available experimental data or to raise up new intriguing and valuable observations about the existence of different, but still viable, modes for molecular recognition and binding even considering the same identical series of inhibitors for thrombin and trypsin proteases.

MATERIALS AND METHODS

Data Set. A number of 72 and 16 well-studied 3-amidinophenylalanine protease inhibitors were used as training and test sets, respectively, to challenge the validity and performance of the MOO method herein proposed.

As previously reported,¹³ Cartesian coordinates for thrombin and trypsin were obtained from 1ETS and 1PPH crystal structures deposited in the Protein Data Bank (PDB).¹⁸ The protonation state of acidic and basic groups was assumed to be as follows: the benzamidine and basic amino functional groups were protonated, whereas aromatic amino functional groups were left uncharged; carboxylic groups were considered to be deprotonated. The experimental inhibitory activities toward the two proteins were expressed as pK_i . Although the proteases share the same catalytic triad formed by Asp102, His57, and Ser195, they present substantial structural differences and rendered data interpretation a difficult task. A higher correlation of affinity data (i.e., between thrombin and trypsin) was associated with fewer residue similarities but still good closeness of 3D structures. Aside from the S2 subsite, where the catalytic triad is located, three main additional subsites in the ligand binding pocket have been highlighted. The first, termed S1, is formed by a deep, narrow hydrophobic cavity at the bottom of which Asp189 forms a salt bridge with positively charged moieties of the inhibitor. All residues are highly conserved, apart from the A190S mutation in trypsin whose pocket is also smaller and slightly more hydrophilic than in thrombin. The second smaller subsite, named D (D stands for distal) basically lined by aromatic residues is also present. The other important

subsite, known as P (P stands for proximal), is particularly evident in thrombin due to the insertion of the loop Tyr60A-Pro60B-Pro60C-Trp60D but absent in trypsin.

Computational Methods. The chemical structures of benzamidine derivatives were taken from the Tripos database¹⁹ available for CoMFA tutorial and subjected to preliminary minimization using the standard *Premim* script released in MacroModel 9.1. Proteins were prepared by adding hydrogen atoms, inserting and optimising missing residues, and removing water and cocrystallized inhibitors. Controls were carefully carried out on the protonation state of polar residues leaning into the active site. Using the *Protein Preparation* module in *Maestro* 7.5,²⁰ light relaxation was performed to optimise hydroxyl and thiol torsions followed by all-atom constrained minimization to relieve steric clashes until the rmsd reached 0.18 Å. Subsequently, the optimized inhibitor structures were docked into the two serine protease by using GOLD.^{21,22} On the basis of the refined crystal structures of thrombin and trypsin, GOLD was set to generate 20 docking poses for each of the 72 and 16 training and test set inhibitors, respectively. Residues within an active radius of 11 Å centered on the α carbon of Gly216 were considered in docking simulations. According to descriptions elsewhere,¹⁶ ligand–receptor complexes were generated by setting physical constraints to favor the occurrence of well-established interactions such as the salt bridge formed by Asp189 and the hydrogen bonds (HB) engaged by Gly216 and Gly219 with the inhibitors. After completing docking simulations, energy scores computed for 20 ligand–enzyme complexes originating from both serine proteases were then annotated into 88×20 sized matrices. Similarly, an identical number of rmsd values were also recorded by measuring distances between corresponding molecular locations occurring in each docking pose compared to designated crystal-based binding hypothesis. The interested reader is referred to elsewhere for a fully detailed procedure description.¹³

Multiojective Optimization. In our previous work based on SOO,^{13,17} the aim was to derive a single QSAR model representing an appropriate balance between best regressions among docking scores with affinities and minimal atomic displacements from properly established crystal-based binding topology. The problem of joining docking and 3D QSAR was effectively reduced to SOO problem through a weighted sum fitness function which involved summing of two different objectives to give a single value. As a result, a single run of SOO led to a single 3D QSAR model only. A further disadvantage of SOO method was the selection of the most appropriate weighting, because it is often not clear how the two different objectives should be calibrated. To overcome these limitations, a novel MOO is herein proposed. MOO is based on the idea of Pareto optimality where a Pareto optimal, or nondominated, solution is one where another solution does not exist that is better than it in both the two objectives. As a result, one solution dominates another if it is either equivalent, or better, in both the objectives and strictly, it is better in at least one objective. This means that two, or more, 3D QSAR models will be equivalent when each of them will better regress docking scores with biological affinities and will worse superimpose to an established crystal-based binding topology or vice versa. In addition, no any other 3D QSAR model must exist which is

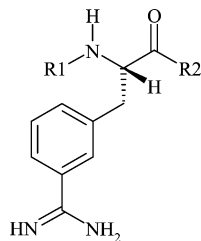


Figure 1. General structure of the examined 3-amidinophenylalanine derivatives.

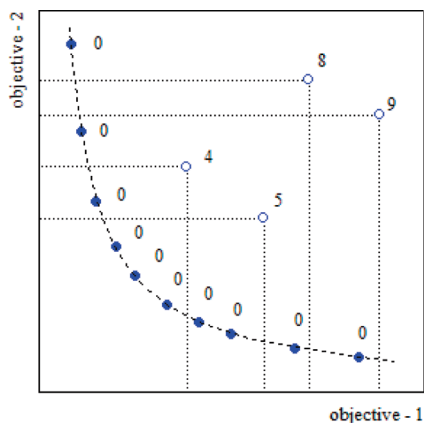


Figure 2. Each point represents a solution to the dual-objective problem. The solid circles are nondominated solutions and fall on the Pareto frontier. Dominated solutions are shown as unfilled circles, and their ranks are assigned according to the number of times they are dominated. Note that the problem requires minimization of both objectives.

better in both the objectives above outlined. In this view, joining molecular docking and 3D QSAR represents a dual-objective problem trading-off the need of selecting alignments having low averaged rmsd value (objective-x) compared to a given binding hypothesis and showing solid correlation between docking scores and pK_i (objective-y). Each potential 3D QSAR model is, thus, assumed to be a point in the search space represented by a pair of properties, for instance the objective-x and the objective-y. A model is nondominated if the square area bounded by the axes and lines drawn parallel with the axes from the point does not include any other point. As indicated in Figure 2, models are therefore ranked according to the number of times they are dominated; thus, nondominated solutions are given rank zero, and the dominated solutions are given ranks according to their position in the graph.

Here, the MOO was included into a genetic algorithm (GA) which is based on the principles of Darwinian evolution and survival of the fittest. Potential solutions are coded as a one-dimensional string of bits whose integer value indicates each specific docking pose. Unlike classical GA, the initial population was created at random through a chaotic blend of bits with the exception of one individual that was, instead, seeded by using the result of our previous SOO application. This initialization procedure proved to be effective in accelerating the entire optimization procedure. Newer generations are formed by propagating offsprings through two basic recombination mechanisms. One is the mutation that causes a random change of separate elements within a chromosome; the other operator is multipoint crossover in which portions of genetic material are taken from each parent and recombined to create new child solutions. At each

iteration, individuals are ranked according to the number of individuals in the population by which they are dominated. Thus, nondominated individuals are assigned rank zero, individuals that are dominated by one individual are assigned rank one, and so on. The fitness of an individual is then determined by its rank, or dominance, with individuals having lower dominance being preferred.

K-Means Clustering. The simultaneous optimization of the two objectives discussed in the present work through the MOO approach resulted in a large family of trade-off models. Despite their equivalence on the basis of the Pareto dominance, a limited number of models can however effectively be used in QSAR practice. Normally, the selected models were biased toward some a priori knowledge or chemical intuition. Model desirability has been sometimes used as a knowledge-based criterion to drive the selection toward those that have been flagged as desirable. Attempts are known to transform model desirability into an objective itself.¹¹ In the present investigation, K-means clustering was used in partitioning the applicability domain of the entire pool of equivalent models and in selecting those local models assumed to better elucidate a way of binding and molecular selectivity of serine protease inhibitors. K-means is a clustering method belonging to unsupervised learning algorithms. Unlike hierarchical clustering,²³ K-means does not create a tree structure but originates a single level of clusters after partitioning model space into K mutually exclusive clusters. Each cluster in the partition is defined by its members and by its centroid that is the point to which the sum of distances from all elements in that cluster is the smallest. Based on an iterative algorithm, K-means minimizes the sum of distances from each object to its cluster centroid over all clusters by moving objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. A measure accounting of the closeness of points in one cluster to points in the neighboring clusters is the silhouette. It ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. The partitioning algorithm operates by calculating intermodel distances after performing principal component analysis of the rmsd values relative to each docking poses included into each equivalent model. Partitions having the highest values of average silhouette were awarded for selecting the number of cluster whose centrotypes designated a subset of representative equivalent molecular alignments for deriving 3D QSAR analyses.

3D QSAR Method. Compounds included in the selected equivalent molecular alignments were analyzed by using GRID²⁴ to obtain molecular interaction fields (MIF) with DRY and OH representative probes typifying hydrophobic and HB acceptor/donor groups, respectively, and to survey the main physicochemical features of the inhibitors. A box size having side and number of grid planes per angstrom equal to 11 Å and to 1, respectively, was used to sample physicochemical space. Analysis of MIFs was conducted with GOLPE,²⁵ an advanced statistics tool used for both variable selection and model validation, where the field matrix was unfolded to produce one-dimension vector variables for each compound. The original matrix with field

values was then pretreated to select unbiased variables by zeroing the positive and negative ones and applying a standard deviation cutoff. The N-level was set on and the weighting procedure, known as block unscaled weights (BUW), was adopted to normalize the importance of each block of variables, particularly the DRY MIFs, by scaling each block separately, whereas the scales of single variables within each block remained unchanged. To obtain 3D QSAR models, partial least-squares (PLS) analyses were performed with the leave-one-out (LOO) cross-validation procedure.²⁶ The optimal number of components (ONC) in the PLS models was considered the one yielding the smallest standard deviation error in prediction (SDEP), given that the increase of q^2 was not higher than 5% when adding another component. The smart region definition (SRD) algorithm was used to group variables (number of seeds = 1500, critical distance = 1.0 Å, collapsing cutoff = 2.0 Å), and two fractional factorial design (FFD) variable selection runs were then executed to obtain the final 3D QSAR model.²⁷ The obtained 3D QSAR models were used to forecast the pK_i values of the 16 inhibitors comprised into the test set (i.e., 73–88, Table 1). All 20 docking poses explored for each test set inhibitor were predicted through extrapolation via 3D QSAR equivalent models. This procedure permitted to explicitly consider pK_i variation at the change of the docking pose, thus, eliminating the bias of selecting just one pose or, even worse, avoiding the user of manually manipulating test set compounds. Test set compounds were, therefore, assigned pK_i values on the basis of *consensus*²⁸ by mean calculation over the 20 docking poses available *per* compound.

RESULTS AND DISCUSSION

Targeting Thrombin and Trypsin through MOO.

Following the strategy outlined above, docking poses generated from IETS refined crystal structure were rescored through MOO. As shown in Figure 3a, a number of 89 solutions were obtained to form the Pareto frontier of equivalent nondominated alignments. Interestingly, it was also observed that the SOO individual seeded into the initial population occurred in the pool of equivalent models as indicated by the solid red circle in Figure 3a. Subsequently, K-means clustering was run on this pool of solutions to subselect from them a bunch of local representative nondominated alignments. Ensuring silhouette point values always higher than zero (Figure 3b), four clusters were returned, and, thus, four centroids were spotted out to designate nondominated models for deriving 3D QSAR.

These selected models were, thus, used as input for GRID.²⁴ DRY and OH interactions were calculated, and PLS regression was performed through GOLPE.²⁵ As shown in Table 2, all the trade-off QSAR models disclosed an outstanding statistics that was validated through intensive randomization tests. For a number of 100 y-scrambling²⁹ runs, it was observed a consistent drop of statistics with the average q^2 being equal to 0.352. A close analysis of the four trade-off 3D QSAR models revealed that 22 out of 72 compounds adopted diverse, though equivalent, molecular orientations in the binding site of the thrombin.

For the ease of presentation, the discussed cases are referred to as the most selective thrombin inhibitors. The highest active thrombin inhibitor **1** adopted three diverse

conformations over the four equivalent alignments. Models A50 and A26 included top-scored docking pose. Models A05 and A09 comprised two very similar conformations having lower scores but improved rmsd; it is worthy saying that model A09 selected a pose already occurred in our previous SOO investigation with this being an indirect proof of the MOO benefit. As illustrated in Figure 4a, inhibitor **1** is, hence, supposed to contact the thrombin binding site via two main modes. The R1 substituent can be directed toward Thr149 enhancing the score value or, alternatively, can be engaged in interaction with Tyr60A of subsite P lowering the rmsd. This latter observation, not occurred in SOO investigation, resulted in a brilliant explanation of molecular selectivity toward thrombin which, unlike trypsin, presented the subsite P. Of great interest is the discussion about inhibitor **6** which adopted (Figure 4b) two diverse conformations. Major differences were noted for R1. It can form HB implicating sulfonyl group and Gly219 and, thus, be located in proximity of methylenic chain of Glu146 in the outer part of the protein. Interestingly, MOO resulted in an additional orientation in which R1 is pointing to the thrombin D subsite, while the interaction with Gly219 was lost. This latter conformation is preferable for its lower rmsd. As illustrated in Figure 4c, inhibitor **8** displayed two fundamental molecular arrangements. One conformation shows the sulfonamidic nitrogen engaged with Glu192 of thrombin through HB reinforced interaction; this can be considered a selective interaction as the Glu192 was mutated to Gln192 in trypsin. In the alternative, but equivalent conformation, inhibitor **8** lost the previous contact to interact with Glu217 of thrombin which is changed in Ser217 in trypsin. In addition, it should be emphasized that such a conformer is remarkably shifted toward the thrombin subsite D. Two different conformations occurred over the four trade-off models for inhibitor **17** as displayed in Figure 4d. On one side, an interaction with Glu192 has been detected; on the other, the substituent R1 faced Trp60A belonging to subsite P which is instead lacking in trypsin.

The application of MOO to trypsin returned a number of 97 solutions located on the Pareto frontier. The analysis of silhouette point values after K-means clustering resulted in four clusters, and, hence, the centroids were subjected to GRID²⁴ and GOLPE²⁵ as already outlined above. Again, it was observed that the SOO seeded model was found in the pool of trade-off models for trypsin too.

As shown in Table 2, a robust statistics was associated with all trade-off QSAR models despite it should kept in mind that the analyzed ligands were primarily conceived as thrombin inhibitors. Randomization tests validated the goodness of the obtained statistics with average q^2 after a number of 100 y-scrambling runs²⁹ equal to 0.340. In the ensemble of nondominated models it was observed that 10 out of 72 compounds adopted diverse, though equivalent, molecular orientations in the trypsin binding site. As already emerged when discussing thrombin binding, the R1 substituent benefited of high conformational freedom. Represented in Figure 5a-c, equivalent solutions for inhibitors **2**, **47**, and **55** disclosed aromatic rings of the R1 substituent rotated around each other of about 180°. Despite their rather diverse location into the trypsin binding site, the interaction between sulfonyl group and Gly219 was always kept with the exception of some weakly active trypsin inhibitors.

Table 1. Substituents at Positions R1 and R2 of the 88 Examined Inhibitors^{a,b}


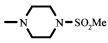
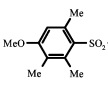
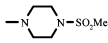
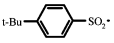
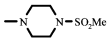
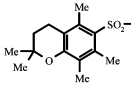

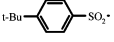

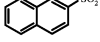



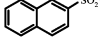

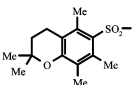
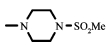
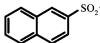


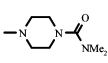
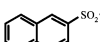



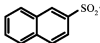
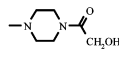


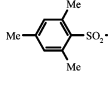


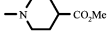
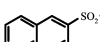


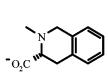
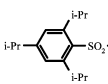


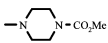
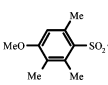



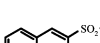
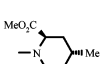

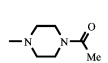
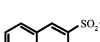

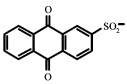

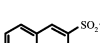
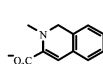
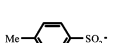

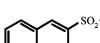

ID	R1	R2	thr	trp	No	R1	R2	thr	trp
1			8.38	6.77	16			7.50	5.75
2			8.37	6.80	17			7.47	6.14
3			8.30	6.70	18			7.43	6.59
4			8.21	6.85	19			7.43	6.66
5			8.13	6.12	20			7.38	6.28
6			8.06	6.77	21			7.38	6.68
7			7.85	6.20	22			7.24	5.96
8			7.80	6.20	23			7.23	5.40
9			7.77	7.44	24			7.19	6.48
10			7.75	6.89	25			7.13	6.16
11			7.72	7.70	26			7.05	6.11
12			7.68	6.26	27			7.02	5.66
13			7.64	6.85	28			6.96	5.85
14			7.59	7.13	29			6.92	5.35
15			7.59	6.28	30			6.92	5.82

Table 1. Continued

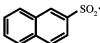



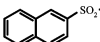



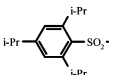



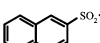



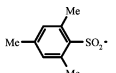



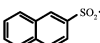



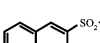



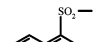

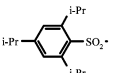

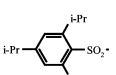



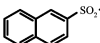



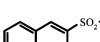

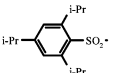

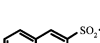

















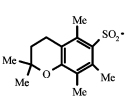

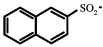

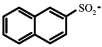

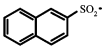

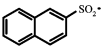

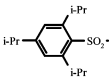

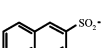



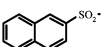

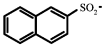
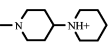
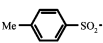

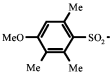

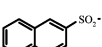



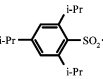

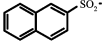

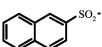

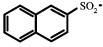

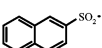
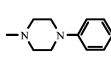
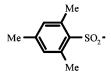

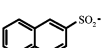

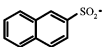

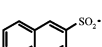
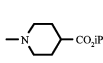
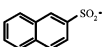

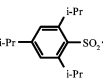



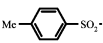

ID	R1	R2	thr	trp	No	R1	R2	thr	trp
31			6.92	5.40	47			6.38	5.68
32			6.82	6.41	48			6.30	6.66
33			6.82	6.57	49			6.29	6.37
34			6.80	6.80	50			6.24	6.24
35			6.75	6.00	51			6.20	6.00
36			6.70	5.92	52			6.18	5.09
37			6.68	7.17	53			6.16	5.92
38			6.64	6.22	54			6.05	6.07
39			6.64	6.10	55			5.96	6.36
40			6.59	6.20	56			5.92	4.85
41			6.55	5.60	57			5.75	6.34
42			6.55	6.92	58			5.68	7.10
43			6.50	5.92	59			5.64	5.10
44			6.47	5.44	60			5.54	4.80
45			6.47	5.92	61			5.51	6.57
46			6.46	5.66	62			5.51	4.50

Table 1. Continued

ID	R1	R2	thr	trp	No	R1	R2	thr	trp
63			5.24	4.60	76			7.52	6.22
64			5.21	4.80	77			7.44	5.89
65			5.14	5.62	78			7.28	6.36
66			4.89	4.54	79			7.16	5.72
67			4.82	6.00	80			6.77	6.15
68			4.77	3.85	81			6.59	6.51
69			4.57	4.54	82			6.55	6.01
70			4.52	3.93	83			6.52	6.80
71			4.46	4.51	84			6.28	7.57
72			4.36	3.00	85			6.28	5.75
73			8.48	6.72	86			6.15	7.64
74			7.89	6.59	87			5.42	4.59
75			7.59	6.50	88			4.75	4.34

^a Taken from reference Bohm and coll. *J. Med. Chem.* **1999**, 42, 458–477. ^b Experimental inhibition data for thrombin (thr) and trypsin (trp) are expressed as p*K_i*.

As well stated in the Kubinyi's paradox, relationships between internal and external predictions cannot be proved since very often high internal predictability may correspond to low external predictability and vice versa.³⁰ On the other hand, the wideness of p*K_i* data is critical to obtain confident predictions normally being a spread equal to 3.0 logarithmic units the lower boundary for deriving statistically significant models.¹⁵ In this view, making a priori assumptions for biasing one single pose or devising

calibration procedures to define weighting coefficients for all the poses are, thus, useless efforts. In the present study, a rather conservative approach was carried out. All docking poses of test set compounds (i.e., **73–88**, Table 1) were individually subjected to each selected 3D QSAR equivalent model to extrapolate their p*K_i* values. Through a consensus by mean, p*K_i* values of each test set compound was, thus, assigned by calculating the average over the 20 docking poses. This procedure was, therefore, repeated

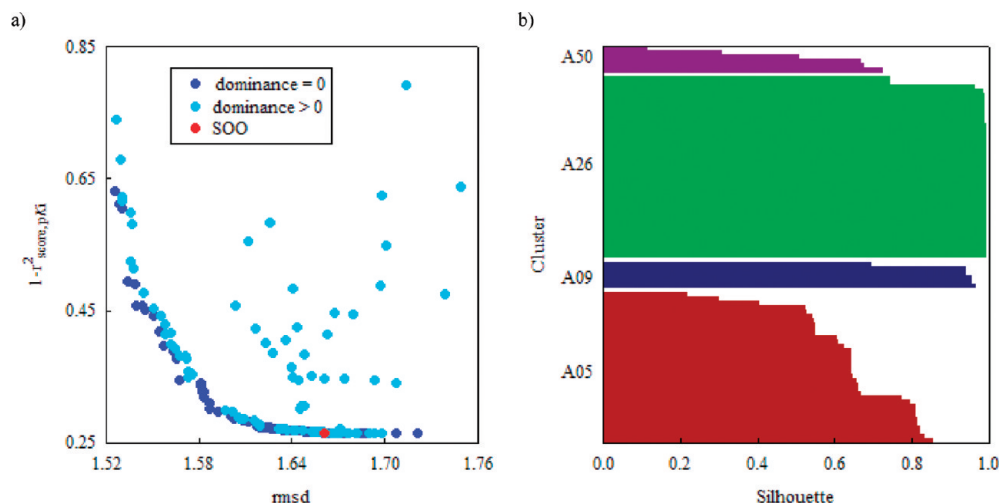


Figure 3. Analyses for thrombin inhibitors: a) blue and cyan solid circles indicate nondominated and dominated models by trading rmsd and r^2_{s,pK_i} objectives; red solid circle indicates nondominated model already emerged in SOO and b) silhouette plot showing the goodness of clustering after applying K-means algorithm to nondominated models.

Table 2. Statistics of Nondominated Thrombin (A) and Trypsin (B) Models Selected through K-Means Clustering^a

centrotype	members	q^2	r^2	SDEC	SDEP
A05	6	0.695	0.831	0.421	0.565
A09	6	0.790	0.827	0.426	0.559
A26	35	0.770	0.867	0.374	0.491
A50	42	0.702	0.888	0.342	0.469
B04	13	0.755	0.861	0.326	0.433
B23	30	0.743	0.850	0.339	0.444
B46	11	0.765	0.838	0.352	0.424
B88	43	0.754	0.855	0.335	0.434

^a Each cluster of nondominated solutions was provided with members and centrotype indicating the number of elements and the molecular alignments effectively used for deriving 3D QSAR. q^2 , r^2 , SDEC, and SDEP represent the leave-one-out cross-validated correlation, the squared correlation coefficient, the standard deviation error in calculation, and the standard deviation error in prediction, respectively. The optimal number of PLS components was always equal to 1.

for each designated equivalent 3D QSAR model and inhibitory values estimated accordingly.

As shown in Table 3, comparisons between experimental and predicted pK_i yielded satisfying values of predictive squared correlation coefficient r^2_{pred} . In particular, r^2_{pred} values spanned from 0.473 to 0.627 for thrombin. This interval comprised the value occurred in SOO investigation (i.e., $r^2_{pred} = 0.554$). The analysis of the trypsin test set resulted in a narrower range of r^2_{pred} values from 0.554 to 0.573 which, however, did not include the SOO model (i.e., $r^2_{pred} = 0.588$).

Previously applied approaches, such as CoMSIA and COMBINE, returned r^2_{pred} values equal to 0.432 and to 0.452 for thrombin and to 0.842 and to 0.426 for trypsin, respectively. However, projections for test set compounds biased one conformation only on the basis of structural similarity to training set compounds. By adopting a consensus by mean in MOO, a medicinal chemist was not required to make any a priori assumption and was instead automatically given a number of viable docking options for better designing new sound inhibitors. Just for knowledge, the retrospective analysis of docking solutions in the search of the magic pose having the smallest pK_i residuals returned values of r^2_{pred} in the range of 0.877–0.907 for thrombin and of 0.768–0.805

for trypsin, respectively. These concepts are illustrated in the plots enclosed in the Supporting Information.

Analyzing Molecular Selectivity. Besides prediction, molecular selectivity is another relevant topic for drug design. It is widely accepted that selective inhibitors are mostly involved in interaction with nonconserved protein residues, whereas nonselective inhibitors generally contact conserved amino acid side chains or backbone. In this view, inhibitors are supposed to be selective whenever rather diverse affinities (i.e., diverse pK_i observed values) are measured for distinct biological targets. Conversely, weak selectivity is associated with inhibitors resulting in similar activity irrespectively of their high or low pK_i values. Of course, optimum would be to have strong difference and high activity toward one target only. To this end, it goes without saying that for explicitly estimating molecular selectivity toward two different targets, at least, a pair of separate models is needed. In the present study, the difference for each compound (i.e., **1–88** of Table 1) between thrombin and trypsin observed pK_i values was, thus, assumed as baseline selectivity measure, the matter to model and to rationalize. As MOO made available four different equivalent models (Table 2) for both thrombin and trypsin, sixteen possible pairwise combinations can, thus, be used to explain the baseline selectivity observed and to gain insights into selective molecular recognition. For the ease of representation, a color map (Figure 6) was adopted to represent molecular selectivity using red- and blue-like tonality to indicate thrombin and trypsin selectivity, respectively, while yellow-like areas were instead indicative of nonselective inhibitors. The top row encoded baseline selectivity with inhibitors sorted on the basis of their experimental selectivity toward thrombin or trypsin. The remaining rows represent the sixteen pairwise combinations formed by relating, in turn, each thrombin with the four trypsin trade-off models in the order listed in Table 2. To facilitate visual inspection, training and test set compounds were displayed separately on the left- and right-hand side, respectively.

Satisfactorily, the large part of training set compounds was properly recalculated by all sixteen pairwise model combinations apart from some nonselective compounds displayed in

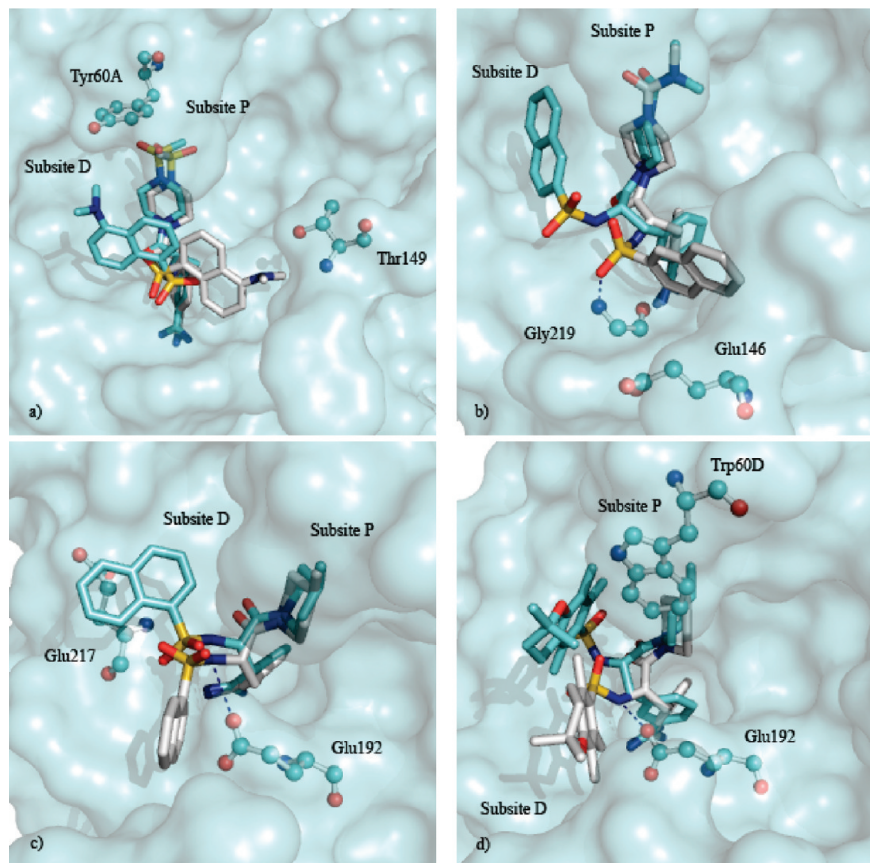


Figure 4. Equivalent docking solutions are colored in cyan and white sticks, and thrombin relevant amino acid residues are rendered as ball and sticks. Aquamarine background surface was rendered from 1ETS thrombin crystal structure. Thrombin inhibitors: a) **1** ($pK_i = 8.38$) with thrombin specific Tyr60A and conserved Thr149 residues; b) **6** ($pK_i = 8.06$) with Gly219 and Glu146 residues; c) **8** ($pK_i = 7.80$) with Glu217 and mutated E192Q residues; d) **17** ($pK_i = 7.47$) with thrombin specific Trp60D and mutated E192Q residues.

the central part of the bar code illustrated in Figure 6. For instance, the affinity of **25** and **26** was always underestimated irrespective of the trade-off trypsin models. This was immediately evident from the low-variant color tonality of the bar codes for each of the four block of thrombin-trypsin model combinations. The likely reason for such a discrepancy could be that these two molecules are close structurally related to compound **16** which, on its own, was only slightly trypsin active compound. Conversely, compounds **66** and **69** were instead overestimated by A09 and A05, respectively. Both of these two alignments selected a different conformation compared to the other equivalent models having instead a single conformation flagging simultaneously the worst docking score and rmsd, in agreement with experimental data. This was likely due to the lack of substituent at the position R1 and to the unexpected localization of R2 in the subsite D.

Selectivity was also analyzed for the test set compounds. As observed, a correct estimation was obtained for most of the thrombin selective compounds. Interestingly, non-selective compounds were also satisfactorily well predicted, while, disappointingly, trypsin selective compounds (i.e., **84** and **86**) were always forecasted as nonselective. Allegedly, compound **86** was not properly explained being that its structure was very similar to that of nonselective compound **49**, while the chemical characteristics of inhibitor **84** were not sufficiently represented in the training set.

Interestingly, each combination of thrombin and trypsin trade-off models worked as a classifier which could assign inhibitors to some specific class. Considering the difference between thrombin and trypsin experimental affinity values (i.e., ΔpK_i), three basic classes were, hence, defined. The first class was represented by thrombin selective compounds that were those having $\Delta pK_i > 1$, the second class was constituted by nonselective compounds showing $-1 < \Delta pK_i < 1$, and the third class was referred to as trypsin selective inhibitors and, thus, resulting in $\Delta pK_i < -1$.

The goodness of each classifier was estimated by using a confusion matrix whose diagonal elements represented inhibitors having selectivity values correctly assigned, while the cross-diagonal elements indicated those misclassified. As a measure of accuracy (AC) in the classification, the following equation was used

$$AC = \frac{a}{a + b + 2c}$$

where a , b , and c indicated compounds correctly predicted, misclassified but without inversion of selectivity and misclassified with inversion of selectivity, respectively. As illustrated in Figure 7, all sixteen classifiers resulting from the combination of the trade-off models resulted in good values of AC in the range between 0.614 and 0.761, and inversion of selectivity was never observed.

Hit Enrichment Experiments. 3D QSAR trade-off thrombin models were finally challenged for their ability to fish

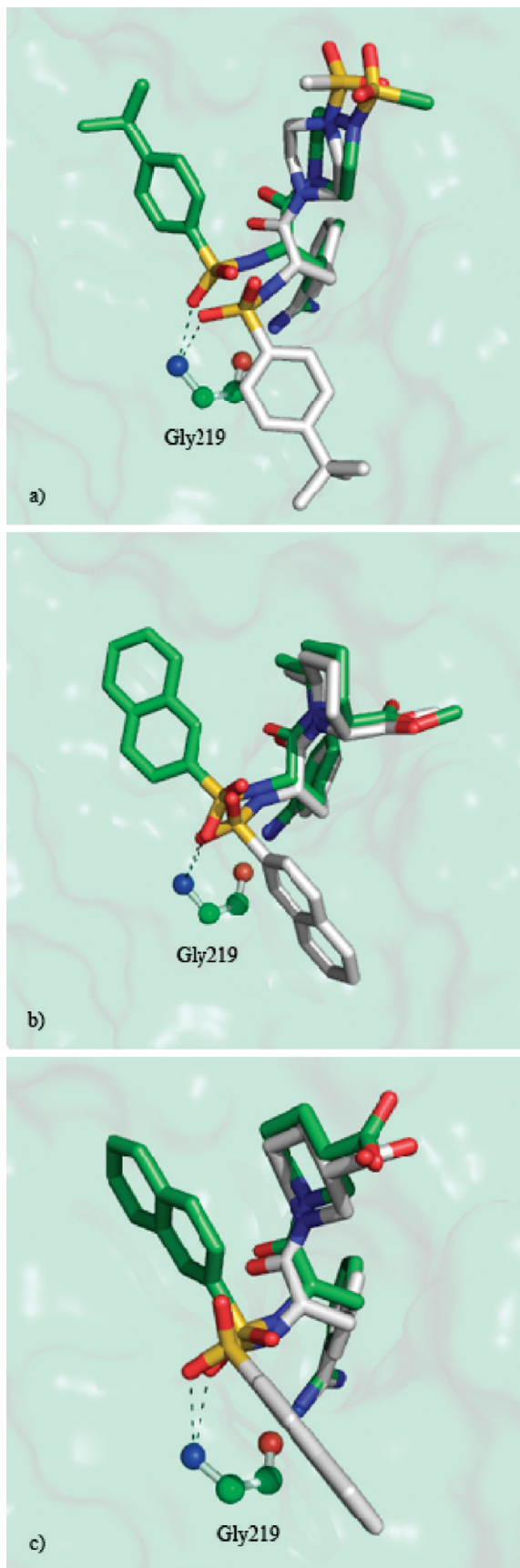


Figure 5. Equivalent docking solutions are colored in green and white sticks, and trypsin relevant amino acid residues are rendered as ball and sticks. Green background surface was rendered from 1PPH trypsin crystal structure, and the occurrence of HB between inhibitors and Gly219 was shown. Trypsin inhibitors: a) **2** ($pK_i = 6.77$); b) **47** ($\Delta pK_i = 5.68$); c) **55** ($pK_i = 6.36$).

Table 3. Predicted pK_i Values of Test Set Compounds Extrapolated on the Basis of Each Trade-off 3D QSAR Model after Consensus by Mean

no.	A05	A09	A26	A50	B04	B23	B46	B88
73	7.04	6.56	7.00	6.81	5.45	5.44	5.45	5.50
74	6.41	6.14	6.42	6.20	4.97	5.02	5.01	5.04
75	7.26	6.86	7.19	7.05	5.59	5.61	5.66	5.66
76	6.87	6.65	6.83	6.79	5.81	5.81	5.83	5.90
77	7.05	6.75	6.99	6.97	5.84	5.86	5.88	5.92
78	6.53	6.32	6.52	6.60	5.55	5.57	5.57	5.62
79	6.67	6.46	6.60	6.57	5.43	5.44	5.45	5.45
80	6.41	6.32	6.45	6.58	5.52	5.53	5.51	5.55
81	6.12	6.22	6.25	6.21	5.87	5.87	5.89	5.92
82	6.28	5.95	6.10	6.02	5.14	5.16	5.21	5.20
83	6.05	6.18	6.09	6.12	5.90	5.93	5.92	5.92
84	6.20	5.97	6.16	6.10	5.96	6.01	6.02	6.01
85	6.49	6.26	6.42	6.36	5.40	5.43	5.44	5.46
86	5.61	5.68	5.69	5.71	5.69	5.73	5.71	5.77
87	5.77	5.61	5.91	5.81	4.42	4.47	4.53	4.47
88	5.91	5.83	6.18	6.18	4.67	4.70	4.68	4.67
r^2_{pred}	0.627	0.584	0.552	0.473	0.554	0.569	0.561	0.573

active hits from a large external benchmarking set of 10500 thrombin ligands available from literature. To minimize random retrieving and to perform an even stringent test, all of the screened benchmarking structures shared the same identical benzamidine core and were, thus, both physically and topologically similar to 3D QSAR training set compounds. The great majority of them (i.e., 10240) were retrieved from a combinatorial library containing 15360 members biased toward thrombin.³¹

The chosen 10240 set of compounds were obtained by reacting a number of 80 aldehydes, 8 amines, and 16 isonitriles according to the Ugi-type three component reaction.

As reported in the original paper,³¹ the isonitrile and aldehyde reactants spanned a broad range of chemical diversity for the inclusion of small aliphatic, aromatic, heteroaromatic, benzylic groups variously decorated with HB donor and acceptor substituents. Incorporating structural pattern suitable for the thrombin arginine binding S1 pocket, the amines were instead known to act as arginine mimetic with affinity toward thrombin in the high micromolar range. The remaining compounds screened were instead recovered from miscellaneous available sources (e.g., the DUD database among the others).³² Hit-inhibitors (183 out of 10500) were designated as only those molecules having pK_i values higher than 6.50 ($K_i < 316$ nM), while all the remaining were considered as decoys (10317 out of 10500). As normally done in standard docking screens, ligand enrichment among top-ranking hits³² was assumed as a key metric for assessing 3D QSAR performances. Maximum predicted pK_i solutions from each trade-off 3D QSAR model and top-scored poses from docking were herein used in the enrichment experiments.

To this end, Figure 9a profiled the percentage of hit-inhibitors as a function of the percentage of the ranked benzamidine pool. Irrespective of the used equivalent model, enrichments obtained through 3D QSAR were always better than that profiled through molecular docking simulations. Similar considerations can be made even when observing the early enrichment indicator EF_1 , quantifying the enrichment factor at 1% of the ranked database.³² Docking based screen resulted in a value of EF_1 equal to 7.10, a number that was overwhelmed by trade-off models ranging EF_1 from 19.12 to 22.40. In other words, it means that, after screening the top 1% of the entire database, docking was able to fish

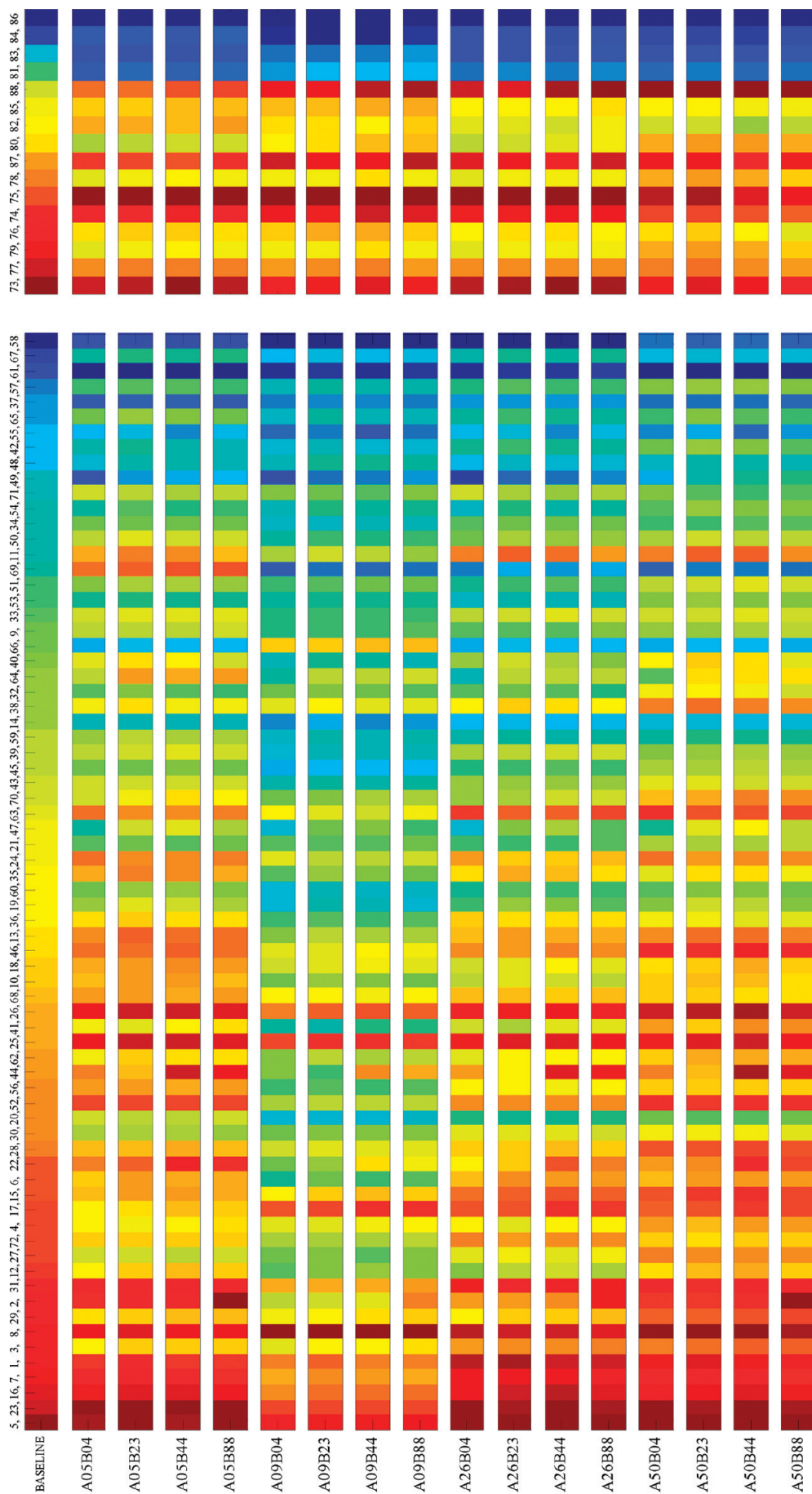


Figure 6. Color bar code. The shift from red- to blue-like colors indicates the change in molecular selectivity from thrombin to trypsin. Top row line (row 1) represents the baseline model that is referred to as difference of thrombin and trypsin experimental activities. Similarly, equivalent models from thrombin and trypsin were pairwise combined to recalculated molecular selectivity (from row 2 to row 17). Left- and right-hand side blocks are referred to as training and test set compounds, respectively. Compounds are sorted in order of descending molecular selectivity as illustrated in row 1.

A05B04		predicted			A05B23		predicted			A05B46		predicted			A05B88		predicted		
		Thr	No	Try			Thr	No	Try			Thr	No	Try			Thr	No	Try
a c t u a l	Thr	17	14	0	a c t u a l	Thr	20	11	0	a c t u a l	Thr	16	15	0	a c t u a l	Thr	15	16	0
	No	15	36	1		No	8	44	0		No	14	38	0		No	10	40	2
	Try	0	3	2		Try	0	2	3		Try	0	3	2		Try	0	4	1
A09B04		predicted			A09B23		predicted			A09B46		predicted			A09B88		predicted		
		Thr	No	Try			Thr	No	Try			Thr	No	Try			Thr	No	Try
a c t u a l	Thr	16	15	0	a c t u a l	Thr	14	17	0	a c t u a l	Thr	17	14	0	a c t u a l	Thr	18	13	0
	No	10	42	0		No	7	45	0		No	8	44	0		No	8	44	0
	Try	0	3	2		Try	0	3	2		Try	0	3	2		Try	0	3	2
A26B04		predicted			A26B23		predicted			A26B46		predicted			A26B88		predicted		
		Thr	No	Try			Thr	No	Try			Thr	No	Try			Thr	No	Try
a c t u a l	Thr	19	12	0	a c t u a l	Thr	19	12	0	a c t u a l	Thr	20	11	0	a c t u a l	Thr	20	11	0
	No	10	41	1		No	8	44	0		No	8	44	0		No	8	44	0
	Try	0	3	2		Try	0	3	2		Try	0	3	2		Try	0	3	2
A50B04		predicted			A50B23		predicted			A50B46		predicted			A50B88		predicted		
		Thr	No	Try			Thr	No	Try			Thr	No	Try			Thr	No	Try
A c t u a l	Thr	21	10	0	a c t u a l	Thr	20	11	0	a c t u a l	Thr	21	10	0	a c t u a l	Thr	20	11	0
	No	12	40	0		No	11	41	0		No	10	42	0		No	8	44	0
	Try	0	4	1		Try	0	4	1		Try	0	4	1		Try	0	4	1

Figure 7. Confusion matrix indicating compound selectivity prediction on the basis of all pairwise trade-off model combinations. Diagonal (light gray) and cross-diagonal (white) elements indicate correct assignment and misclassification, respectively.

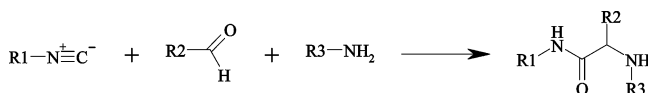


Figure 8. Ugy-type three-component reaction involving 16 isonitriles, 80 aldehydes, and 8 amines to generate a combinatorial library of 10240 product compounds.

the 12.37% of hit-inhibitors, while trade-off models outperformed with recoveries from 33.32% to 39.04%.

Interestingly, it was observed that model A09 returned both better enrichments (i.e., dominant green solid line in Figure 9a) and higher QSAR statistics ($q^2 = 0.790$, $r^2 = 0.827$ in Table 2). This trend held true even when screening small percentages (i.e., less than 30%) of ranked database. This concept became even clearer by profiling the receiver operator characteristic (ROC) curve which related the term [1 - specificity] (i.e., % of selected decoys) vs sensitivity (i.e., % of selected hit-inhibitors). As illustrated in Figure

9b, our 3D QSAR equivalent models denoted a higher sensitivity in comparison to docking. As a further confirmation, a larger area under curve (AUC) was obtained for 3D QSAR trade-off models ($0.660 \leq \text{AUC} \leq 0.744$) than for docking (AUC = 0.597) virtual screening. As to some extent expected, visual inspection of hits revealed that 3D QSAR was particularly able to correctly rank true active molecules with remarkable structural similarity to training set. In this perspective, our 3D QSAR trade-off models would serve as a useful support for designing focused combinatorial libraries and as a diagnostic protocol for their successful screening.

CONCLUSIONS

The MOO approach demonstrated to enlarge standard QSAR boundaries with the promise of inaugurating a new way of thinking about QSAR. And in fact, MOO enabled the exploration of the trade-off between two key-point

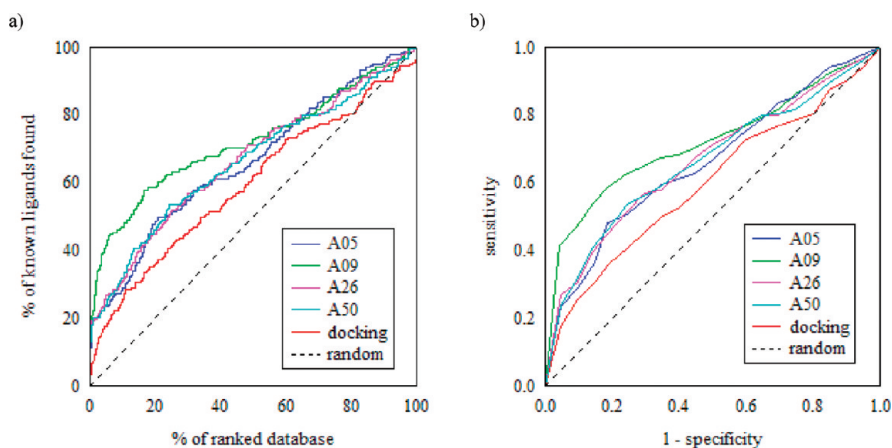


Figure 9. (a) Enrichment plot and (b) ROC curve profiling hit retrieving based on thrombin trade-off 3D QSAR models, docking virtual screening and random selection.

objectives in molecular design. One is the need of explaining the variance of biological affinity through linear regressions; the other is the reproduction of the binding mode observed in X-ray solved protein structures. Without requiring any a priori calibration and weighting scheme, MOO is simply based on a function of dominance to rank potential optimal models. Unlike SOO, the pragmatic effect of MOO is, thus, the obtaining of a family of equivalent models representing different compromises to a given problem. The medicinal chemist is, therefore, given the chance of selecting an appropriate solution better engaging his intuition and feeling by fishing in a pool of a number of equally valid alternative 3D QSAR models. It goes without saying that MOO offered a number of explicit and valuable advantages compared to standard QSAR methods. It was in fact demonstrated that MOO enabled the detection of the occurrence of equivalent but different, binding conformations capturing diverse available potential sites of interaction within the protein binding site. In addition, MOO was very effective in interpreting and quantifying enzyme selectivity.³³ In particular, MOO offered a panel of solutions on the basis of all pairwise trade-off model combinations, and, unequivocally, inversion of selectivity was never occurring in the case of strong selective inhibitors. Finally, trade-off 3D QSAR models disclosed an outstanding sensitivity, improving docking screens, in recovering hits from a large chemical collection of decoys. The promise of a successful design of focused chemical libraries is the natural perspective of this study.

ACKNOWLEDGMENT

Credits are due to Dr. Giuseppe Mangiatori, Dr. Lydia Siragusa, and Dr. Paola Tedeschi for helpful discussion. The authors gratefully acknowledge the financial support from European Commission ("CancerGrid" STREP project, FP VI, Contract LSHC-CT-2006-03755) and MIUR (Rome, Italy; PNR project RBNE01F5WT).

Supporting Information Available: External predictions based on trade-off models derived through MOO—thrombin test set and trypsin test set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- Gillet, V. J.; Nicolotti, O. Evaluation of reactant-based and product-based approaches to the design of combinatorial libraries. *Perspect. Drug Discovery Des.* **2000**, *20*, 265–287.
- Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
- Cottrell, S. J.; Gillet, V. J.; Taylor, R. Incorporating partial matches within multi-objective pharmacophore identification. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 735–749.
- Vainio, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.
- Dey, F.; Caffisch, A. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679–690.
- Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving interpretable structure-activity relationship models. 2. Using multiobjective optimization to derive multiple models. *J. Chem. Inf. Model.* **2008**, *48*, 1558–1570.
- Nicolotti, O.; Gillet, V. J.; Fleming, P.; Green, D. A Novel approach to deriving accurate and chemically intuitive QSAR models. *J. Med. Chem.* **2002**, *45*, 5069–5080.
- Nicolotti, O.; Gillet, V. J.; Fleming, P.; Green, D. Blackwell Publishing 003, 264. A multi-objective approach to deriving QSAR models. In Ford, M.; Livingstone, D.; Dearden, J.; Van De Waterbeemd, H. *Designing Drug And Crop Protectants: Processes, Problems and Solutions*; Blackwell Publishing: Massachusetts, U.S.A., ISBN: 1-4051-2516-0.
- Nicolotti, O.; Carotti, A. QSAR and QSPR studies of a highly structured physicochemical domain. *J. Chem. Inf. Model.* **2006**, *46*, 264–276.
- Stockfisch, T. P. J. Partially unified multiple property recursive partitioning (PUMP-RP): a new method for predicting and understanding drug selectivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1608–1613.
- Cruz-Monteagudo, M.; Borges, F.; Cordeiro, M. N. Desirability-based multiobjective optimization for global QSAR studies: application to the design of novel NSAIDs with improved analgesic, antiinflammatory, and ulcerogenic profiles. *J. Comput. Chem.* **2008**, *29*, 2445–2459.
- Nicolaou, C. A.; Brown, N.; Pattichis, C. S. Molecular optimization using multi-objective methods. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 316–324.
- Nicolotti, O.; Miscioscia, T. F.; Carotti, A.; Leonetti, F.; Carotti, A. An integrated approach to ligand- and structure-based drug design: development and application to a series of serine protease inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 1211–1226.
- Dorfman, R. J.; Smith, K. M.; Masek, B. B.; Clark, R. D. A knowledge-based approach to generating diverse but energetically representative ensembles of ligand conformers. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 681–691.
- Böhm, M.; Stürzebecher, 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*, 458–477.
- Murcia, M.; Morreale, A.; Ortiz, A. R. Comparative binding energy analysis considering multiple receptors: a step toward 3D-QSAR models for multiple targets. *J. Med. Chem.* **2006**, *49*, 6241–6253.
- Nicolotti, O.; Giangreco, I.; Miscioscia, T. F.; Carotti, A. Investigating enzyme selectivity and hit enrichment by automatically interfacing ligand- and structure-based molecular design. *QSAR Comb. Sci.* **2009**, *28*, 861–864.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhatt, N.; Shindyalov, I. N.; Bourne, P. E. The protein Data Bank. *Nucleic Acid Res.* **2000**, *28*, 235–242.
- SYBYL, version 7.1; Tripos Inc.: 1699 South Hanley Road, St. Louis, MO 63144, 2007.
- Maestro, version 7.5.112; Schröedinger, LLC: New York, 2006.
- Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R. L.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
- Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52*, 609–623.
- Jain, A. K.; Murty, M. N.; Flynn, P. J. Data Clustering: a Review. *ACM Comput. Surv.* **1999**, *31*, 264–323.
- Goodford, P. J. A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
- Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 131–150.
- Pastor, M.; Cruciani, G.; Clementi, S. Smart region definition: a new way to improve the predictive ability and interpretability of three-dimensional quantitative structure-activity relationships. *J. Med. Chem.* **1997**, *40*, 1455–1464.
- Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
- Tropsha, A.; Gramatica, P.; Gombar, V. J. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
- Doweyko, A. 3D-QSAR illusions. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587–596.
- Ilgen, K.; Enderle, T.; Broger, C.; Weber, L. Simulated molecular evolution in a full combinatorial library. *Chem. Biol.* **2000**, *7*, 433–441.
- Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
- Nicolotti, O.; Miscioscia, T. F.; Leonetti, F.; Muncipinto, G.; Carotti, A. Screening of matrix metalloproteinases available from the PDB: insights into biological functions, domain organization and zinc binding groups. *J. Chem. Inf. Model.* **2007**, *47*, 2439–2448.