Bioactive Conformational Biasing: A New Method for Focusing Conformational Ensembles on Bioactive-Like Conformers

Boaz Musafia† and Hanoch Senderowitz*

Epix Pharmaceuticals Ltd., 3 Hayetzira St., Ramat Gan 52521, Israel

Received May 6, 2009

Computational approaches that rely on ligand-based information for lead discovery and optimization are often required to spend considerable resources analyzing compounds with large conformational ensembles. In order to reduce such efforts, we have developed a new filtration tool which reduces the total number of ligand conformations while retaining in the final set a reasonable number of conformations that are similar (rmsd ≤ 1 Å) to those observed in ligand−protein cocrystals (bioactive-like conformations). Our tool consists of the following steps: (1) Prefiltration aimed at removing ligands for which the probability of finding bioactive-like conformations is low. (2) Filtration based on a unique combination of two-/three-dimensional ligand properties. Within this paradigm, a filtration model is defined by its workflow and by the identity of the specific descriptors used for filtration. Thus, we developed multiple models based on a training set of 47 drug compounds and tested their performance on an independent test set of 24 drug compounds. For test set compounds after prefiltration, our best models have a success rate of ∼80% and were able to reduce the total number of conformations by 36% while maintaining a sufficiently large number of bioactive-like conformations and slightly increasing their proportion in the filtered ensemble. We were also able to reduce by 39% the number of conformations that are remote (rmsd > 2.5 Å) from the bioactive conformer (nonbioactive conformations). In accord with previous reports, prefiltration is shown to have a major effect on model performance. The role and performance of specific descriptors as filters is discussed in some detail, and future directions are proposed.

INTRODUCTION

In silico or virtual screening has become a common practice in current computer-aided drug discovery efforts.[1−5] The method is used to reduce the time and the cost involved in identifying hits which bind to biological targets of pharmaceutical interest by computationally screening thousands to millions of virtual (most are often commercially or in-house available) compounds for their binding and sometimes their ADME/T (absorption, distribution, metabolism, excretion, toxicity) properties.[4,6] The resulting virtual hits are then tested in vitro for their bioactivity. Hits obtained through screening campaigns, either in silico or in vitro, require further work to optimize them into early drug candidates (EDCs). Similar to screening, this "lead optimization" phase can also greatly benefit from in silico approaches.

The methods most often used in these fields could be broadly classified into structure- and ligand-based approaches. Structure-based in silico screening utilizes the three-dimensional (3D) structure of a target protein in order to dock a set of virtual compounds in an attempt to identify those that best interact with its binding site.[7] In this way, the protein's binding site effectively constrains the conformational space available to the ligand, thereby increasing the probability of finding bioactive-like conformations. The success of this approach, therefore, critically relies on the availability of reliable 3D structures of target proteins which could be obtained either experimentally[8] or computationally.

In the absence of a reliable 3D structure of the target protein, ligand-based approaches could be used. In such cases, known active compounds are employed as templates to screen against a set of virtual compounds. Common methods in this field include two-dimensional (2D) and 3D database searches, pharmacophore queries, and 2D and 3D QSAR.[9,10] The success of ligand-based in silico screening requires the ability to reproduce the bioactive conformers, namely, those conformers which are expected to dominate the protein−ligand complex ensemble of both known (template) and virtual (target) ligands.

In the absence of a 3D structure of the protein's active site to direct the generation of bioactive-like conformers, these could, in principle, be obtained through conformational search simulations performed either in vacuo or in solvent models.[11−13] Several groups have examined the ability of conformational search methods to produce cocrystal-like conformers using methods such as Catalyst,[14−19] MacroModel,[14−16,18] Omega,[14,15,18,20,21] MOE,[18,19] SPE,[18,22] Rubicon,[18] FROG,[23] IMC,[16] Confort,[14] and Flo99.[14] Successes in these studies were usually defined as the ability to obtain at least one conformation with a root-mean-square deviation (rmsd) of 0.6 Å or less from the crystal structure. However, such conformations seldom ranked among the lowest energy ones, supporting the hypothesis that neither the in vacuo nor the solution structure of an isolated ligand is expected to resemble the bioactive conformer. This is consistent with the published analyses of the 3D structure of ligands complexed with proteins,[14−18,24−26] which revealed the following: (1) Bioactive conformers tend to be more

* Corresponding author. E-mail: hsenderowitz@gmail.com.
† Current address: bmusafia@gmail.com.

10.1021/ci900163t CCC: $40.75
Published on Web 10/14/2009

elongated than those of unbound minimized ones; (2) The strain energies of bioactive conformers are higher than those of the global minima; (3) In many cases, bioactive conformers do not correspond to a (local) energy minimum on the potential energy surface of the unbound compound; (4) When a ligand interacts with different targets, it adopts different bioactive conformers. Thus the term "bioactive" should only be viewed in the context of a specific biological target. An additional limitation of current conformational search methods is that many of these were developed with conformational diversity in mind. Consequently, bioactive conformations, when generated, are often accompanied by many conformations which are remote from the bioactive one, thereby introducing noise into the system.

While the area of structure- and ligand-based drug design is scattered with many problems, the current work was motivated by the challenges and the frustrations of spending significant time and resources on the analysis of nonrelevant ligand conformations during lead discovery and optimization campaigns. With this in mind, we developed a new conformational ensemble filtration method aimed at reducing the total number of conformations while maintaining a sufficiently large number of bioactive-like ("good") conformations (defined by a rmsd $\leq$ 1 Å with respect to the crystal structure) to allow for the usage of ligand-based approaches. We have also sought to increase the proportion of bioactive-like conformations or alternatively to reduce the proportion of nonbioactive ("bad") conformations (defined by a rmsd $>$ 2.5 Å with respect to the crystal structure) in the filtered ensemble. Our method starts by identifying and removing ligands for which the probability of finding bioactive-like conformers is small. All the remaining ligands are subjected to a conformational search procedure, and the resulting conformational ensembles are filtered based on 2D and 3D molecular descriptors to produce a subset of the input ensemble enriched by bioactive-like conformations and impoverished by nonbioactive conformations. Indeed, a unique characteristic of this method is the explicit reference to such nonbioactive conformations, an aspect that has not been explicitly dealt with before.

Compounds for the present study were selected from the SuperDrug database[27] (see Supporting Information, Table 1−3). This database contains a chemically diverse set of 183 drugs whose complexes, with their respective biological targets, were crystallographically determined at a resolution of 1.2−3.45 Å. Our final selection of 71 complexes emphasize diversity in both ligand and target spaces with 52 unique proteins, 40 of which (77%) having only a single occurrence (see Supporting Information, Table 4). Reliance on real drugs for model development and validation as well as maintaining high ligand/protein diversity are unique to this work and, in our view, an improvement over previous studies in the field.[14,16]

Multiple models were developed from a diverse set of 47 drug compounds (the training set). These models differed from one another by the identity of the molecular descriptors and by their corresponding cutoff values (see Methods Section). The best performing models were subsequently tested on an independent test set of 24 diverse drug compounds.

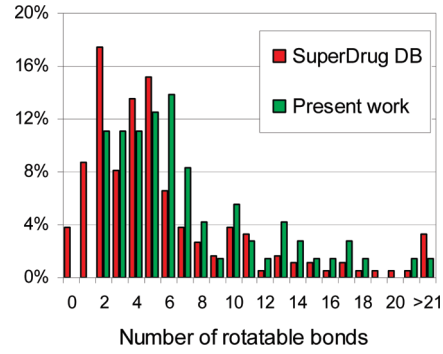

Figure 1. A comparison of the ligand distribution with respect to the number of rotatable bonds in the entire SuperDrug database and in the subset used in the present work.

METHODS

Data Set Selection. A total of 71 compounds were selected from the SuperDrug database based on several considerations:

Number of Rotatable Bonds (RotBond). Our selection uniformly covers a range of two to seven rotatable bonds. Compounds with 8−14 rotatable bonds are represented to a lesser extent, and a few compounds with up to 22 rotatable bonds are also included (see Figure 1 and Supporting Information Table 1). This is in contrast with some other publications in this field that have focused on compounds with around six rotatable bonds.[19]

Protein Resolution. Protein resolution was targeted to less than 2.5 Å, as shown in the Supporting Information, Table 1 (except form complexes with lovastatin, deravirdine and ibuprofen, which had 2.6 Å resolution). This criteria was based on the work of Bostrom et al. 2001,[14] who pointed out that the positional error is expected to be about one-sixth of the resolution. Thus, the expected experimental error is roughly 0.45 Å, which is in line with a 1 Å rmsd threshold deemed appropriate for lead optimization purposes.

Ligand and Protein Diversity. Care was taken to select a diverse set of ligands that were cocrystallized with different proteins. Ligand diversity was assessed based on the number of rotatable bonds and on a visual inspection of the SuperDrug database.

Training Set and Test Set Selection. The prioritized subset of 71 ligands was sorted first by the number of rotatable bonds, then by the number of heavy atoms, and finally by the protein resolution. It was then divided into training and test sets containing 47 and 24 compounds, respectively; (a 2:1 ratio) by allocating representative compound along the sorted list to the test set. This methodology ensured that both the training and test sets had similar profiles in terms of the above three sorting parameters.

Generation of Conformations. A list of ligand−protein crystal complexes was retrieved from the SuperDrug database, and the corresponding ligands were extracted from the PDB. In order to avoid any potential bias resulting from crystal packing forces, all ligands were minimized prior to conformation generation. Minimization was performed using the smart minimization protocol implemented in Discovery Studio[28] using 500 minimization steps, a MMFF force field, and a distance dependent dielectric constant of 1.

Ligand conformations were generated using two conformational search methods:

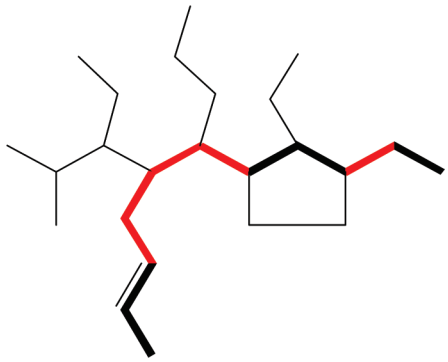

Figure 2. M-PRoB is calculated by counting the number of rotatable bonds (i.e., excluding double bonds, aromatic bonds, bonds within ring systems, and terminal bonds having a carbon atom connected to three identical atoms) along the maximal bonded path. In the figure, the maximal path is bolded, and the relevant rotatable bonds are colored in red.

(1) The Monte Carlo multiple minima (MCMM) algorithm[29] as implemented in Schrödinger's MacroModel.[30,31] The default setup was used with 5 000 MC steps and with the removal of enantiomers. New conformations were identified on the basis of an all-atom (except nonpolar hydrogen atoms) comparison with a maximal distance threshold of 0.25 Å. Minimization was performed with the TNCG optimizer, using a MMFF force field and a distance dependent dielectric constant of 1, and terminated after 500 steps or once the gradient fell below a threshold of 0.05 kJ/mol.

(2) The poling algorithm[32] as implemented in Accelry's Discovery Studio.[28] The following setup was used: conformational method = best, max conformations = 10 000, pole value = 110, energy threshold = 20 kcal/mol, and discard existing conformation option activated.

It is important to note that the number of generated conformations by both methods was not kept constant but rather controlled by the energetic cutoff and by conformational diversity (a rmsd threshold was used to distinguish between different conformations). As common in conformational search procedures, a high upper limit on the number of Monte Carlo (MacroModel) or Polling (Catalyst) cycles was set (5 000 and 10 000 respectively).

Development of a New Descriptor. A new molecular descriptor was developed based on two widely used descriptors, namely, the total number of rotatable bonds (RotBond) and the maximal bonded path along the compound (MaxPath). The new descriptor termed max-path-rotatable-bond (M-PRoB) counts the number of rotatable bonds along the maximal path only and, consequently, is expected to provide a better description of the overall molecular flexibility than RotBond (see Figure 2). The new descriptor was calculated with a Discovery Studio protocol that was developed with assistance from Accelrys, Inc.

Bioactive-Like and Nonbioactive Conformations. For the purpose of model development, bioactive-like conformations were defined as having a rmsd ≤ 1 Å with respect to the cocrystal structure (which we assume to represent the bioactive conformer), whereas nonbioactive conformations were defined as having a rmsd > 2.5 Å with respect to the cocrystal structure. The rmsd between the heavy atoms of each conformation and the crystal structure was calculated using Schrödinger's Maestro.[30]

Descriptors Calculation. Fifteen 2D descriptors (M-PRoB, ALogP, MW, JX, JY, Wiener, flexibility index (PHI), MaxPath, number of bonds, hydrogen atoms, rotatable bonds, rings, aromatic rings, and hydrogen-bond (H-bond) acceptors and donors) were calculated for each of the ligands comprising the training and test sets and 10 3D descriptors (maximal distance between MaxPath atoms, energy, strain energy, radius of gyration, Jurs SASA, principle moment of inertia magnitude (PMI-mag), shadow nu, and shadow *x*-, *y*-, and *z*-lengths) for each of their conformations. All descriptors were calculated using Accelrys' Discovery Studio.[28] See Supporting Information for a short description of each of the descriptors.

In order to be able to compare the performances of different 3D descriptors, we have normalized, on a per compound basis, the descriptor values based on the approach of Diller and Merz,[26] according to eq 1, and divided each descriptor into 11 bins:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

Where x is the descriptor value and μ and σ are, respectively, its mean and standard deviation over all conformations of a single compound.

The normalized values (Z) were divided into 11 bins (bin-0 to -10) based on their standard deviation relative to the mean (bin-0: $-4.5 \geq Z$; bin-1: $-4.5 < Z \leq -3.5$; bin-2: $-3.5 < Z \leq -2.5$; bin-3: $-2.5 < Z \leq -1.5$; bin-4: $-1.5 < Z \leq -0.5$; bin-5: $-0.5 < Z \leq 0.5$; bin-6: $0.5 < Z \leq 1.5$; bin-7: $1.5 < Z \leq 2.5$; bin-8: $2.5 < Z \leq 3.5$; bin-9: $3.5 < Z \leq 4.5$; and bin-10: $Z > 4.5$).

Model Generation. *Basic Workflow.* In the present study, a filtration model is completely defined by its workflow and by the identity of 2D and 3D descriptors used at each stage. Figure 3 presents the basic workflow of our model, which consists of the following steps:

(1) Prefiltration using a preselected 2D descriptor with the aim of identifying compounds for which the probability of identifying bioactive-like conformers is low. For this purpose, we have examined two descriptors as prefilters, namely, RotBond and the newly developed descriptor M-PRoB.

(2) Division of the remaining compounds into two groups based on 2D descriptors. Division into a larger number of groups is possible but may lead to overfitting.

(3) Within each group, application of a (potentially different) 3D descriptor to identify those bin(s) which are most enriched by bioactive-like conformations.

Selection of the Best 2D-3D-3D Descriptors Combination. Given the above-described workflow and the identity of the prefiltering descriptor, the performance of each model only depends on the identity of 2D and 3D descriptors and on their corresponding cutoff values.

2D Descriptors. The training set was first sorted according to the nominal values of the 15 2D descriptors (Figure 3 in pink). Next, the compounds were split into two groups subjected to each group containing no less than five compounds (a total of 28 splits for each descriptor for the 37 compounds of the training set after prefiltration).

3D Descriptors. For each of the two groups resulting from each 2D split (yellow and blue in Figure 3), all 10 3D descriptors (calculated for every conformation) were evalu-

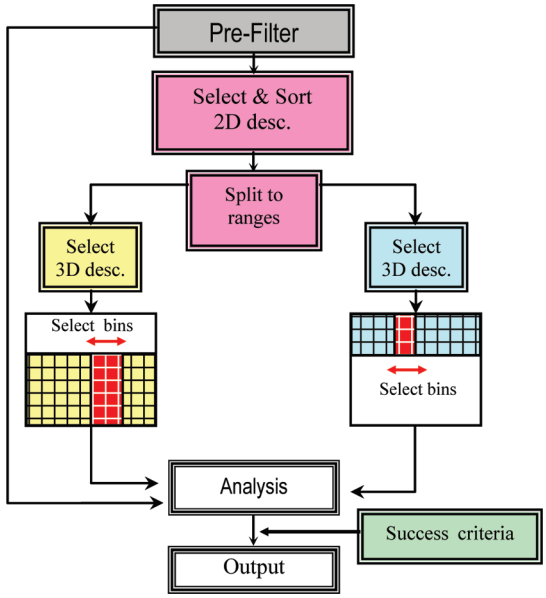

Figure 3. A schematic representation of the model development workflow. Two 2D descriptors were considered for prefiltration, RotBond and M-PRoB (see text for more details). All fifteen 2D descriptors were considered for splitting the remaining compounds into two groups, and all possible splits resulting in both group sizes ≥5 were attempted. For each such split, all ten normalized 3D descriptors were considered for focusing on the output conformations' subset which best satisfies the success criteria and the goodness-of-hit list measure. A total of ~6 × 10^6 2D-3D-3D combinations (i.e., models) were, therefore, attempted, and those that best performed on the training set were subsequently tested on the independent test set.

ated to see whether bioactive-like conformations tend to cluster in specific normalized bins. The performance of each such 2D-3D-3D combination (a model) was evaluated according to the following scheme:

(1) For each ligand, we have determined whether, following filtration, a sufficiently large number of bioactive-like conformations was retained (see Table 1).

(2) Models, for which criterion (1) above was met for at least 35 out of 37 training set compounds (after prefiltration), were evaluated for their average (over all training set compounds) enrichment of bioactive-like conformations and for their average impoverishment of nonbioactive conformations. The enrichment factor was calculated according to eq 2:[33]

$$\text{enrichment} = \frac{TP/n}{A/N} \quad (2)$$

Where N and A are, respectively, the total number of conformations and the total number of bioactive-like conformations for the ligand and n and TP (true positive) are, respectively, the total number of conformations and of bioactive-like conformations in the selected subset of the ligand. The impoverishment factor was calculated in a similar way using the number of nonbioactive conformations in the input conformational ensemble and in the selected subset.

Only models with an average bioactive-like enrichment factor ≥1.2 and nonbioactive impoverishment factor ≤0.4 were further considered. Although in principle, we would have been satisfied with models meeting either one of the enrichment/impoverishment criteria, imposing the requirement that each model would meet both helped in reducing the final set of plausible models. We also reasoned that such models would better perform on the test set. While the exact threshold values listed above are not supported by strong theoretical arguments, they do reflect specific requirements derived from multiple ligand-based lead discovery and optimization campaigns.

Table 1. Success Criterion 1: Number of Bioactive-Like Conformations Needed to Be Retained in the Final Conformational Ensemble as a Function of the Number of Bioactive-Like Conformations Present in the Input Conformational Ensemble

no. of bioactive-like conformations prefiltration	no. of bioactive-like conformation retained postfiltration
1−2	all
3−20	≥50%
>20	at least 10

The models that met these success criteria were further ranked according to the average goodness-of-hit list measure (eq 3),[33] which provides a good balance between specificity, selectivity, and yield:

$$\text{goodness-of-hit list} = \left(\frac{3}{4} \times Ya + \frac{1}{4} \times Se\right) \times Sp \quad (3)$$

The bioactive-like yield (Ya) of the model is given by eq 4:

$$Ya = \frac{TP}{n} \quad (4)$$

Model selectivity (Se) is given by eq 5:

$$Se = \frac{TP}{A} \quad (5)$$

Model specificity (Sp) is given by eq 6:

$$Sp = \frac{TN}{N - A} \quad (6)$$

Where all parameters have their meaning as defined above, and TN (true negative) is the number of bioactive-like conformations that were not included in the final subset.

Note that, in those cases where all conformations were classified as bioactive-like or where no bioactive-like conformations were identified, the enrichment cannot be defined, and consequently, the enrichment criterion could not be met. Similarly, when no nonbioactive conformations were identified, the impoverishment criterion could not be met.

Due to the large number of possible models (~6 × 10^6 2D-3D-3D combinations), model generation and evaluation were automated through a Perl script. The best performing models were subsequently evaluated on the test set. Model application to a specific test set compound was deemed successful if, following filtration, the remaining conformational ensemble contained a sufficiently large number of bioactive-like conformations (Table 1) and had an enrichment factor ≥1.2 or an impoverishment factor ≤0.4.

RESULTS AND DISCUSSION

In this work, we present a new method for focusing conformational ensembles on bioactive-like conformers. Such

Table 2. Number of Ligands[a]

a. Number of Ligands with at Least One Bioactive-Like Conformation (rmsd ≤ 1 Å) as a Function of M-PRoB

	M-PRoB 1−3	M-PRoB 4−5	M-PRoB 6−7	M-PRoB 8−9	M-PRoB 10−18
MacroModel	16	7	9	0	3
DS	16	8	3	1	0
total number of ligands	16	10	11	5	5

b. Number of Ligands Having More than 50% of Their Conformations as Nonbioactive (rmsd > 2.5 Å) as a Function of M-PRoB

	M-PRoB 1−3	M-PRoB 4−5	M-PRoB 6−7	M-PRoB 8−9	M-PRoB 10−18
MacroModel	0	1	2	5	5
DS	0	1	2	4	5
total number of ligands	16	10	11	5	5

c. Number of Ligands with at Least One Bioactive-Like Conformation (rmsd ≤ 1 Å) as a Function of the Number of the Rotatable Bonds

	RotBond 2−3	RotBond 4−5	RotBond 6−8	RotBond 9−12	RotBond 13−
MacroModel	11	8	13	0	3
DS	11	8	7	2	0
total number of ligands	11	8	15	7	6

d. Number of Ligands Having More than 50% of Their Conformations as Nonbioactive (rmsd > 2.5 Å) as a Function of the Number of Rotatable Bonds

	RotBond 2−3	RotBond 4−5	RotBond 6−8	RotBond 9−12	RotBond 13−
MacroModel	0	0	1	6	6
DS	0	0	1	5	6
total number of ligands	11	8	15	7	6

[a] The total number of ligands within the M-PRoB range, whether they have bioactive-like conformation or not, is presented at the bottom row of the table. The results obtained with both MacroModel and Discovery Studio (DS) are presented.

ensembles, while often containing the bioactive-like conformers, are almost always contaminated by many other conformations (in our data set most have an rmsd between 1 and 2.5 Å from the crystal structure, but quite a few have an rmsd > 2.5 Å), thereby introducing noise and bias into the system. Filtering out such conformations, while retaining the bioactive-like ones (rmsd ≤1 Å, see below), is, therefore, expected to greatly enhance the success and the efficiency of ligand-based methods. With this in mind, the criteria for a successful filtration model were defined as follows: (1) reduction of the total number of the conformations in the final set while maintaining a sufficiently large number of bioactive-like conformations, (2) enrichment of the final set by bioactive-like conformations, and (3) removal of as many nonbioactive conformations as possible. Through visual inspection of many conformational ensembles, we observed that conformations deviating by a rmsd ≤ 1 Å from the crystal structure largely resemble it, while conformations with a rmsd > 2.5 Å with respect to the crystal structure are clearly different from it. Thus, rmsd thresholds of ≤1, and >2.5 Å were used to define "bioactive-like" and "nonbioactive" conformations, respectively, where the second value has some support in the literature.[32,33]

Two conformational search algorithms were considered in this work, namely, the Monte Carlo multiple minimum (MCMM) algorithm[29] implemented in MacroModel[30] and the poling algorithm implemented in Discovery Studio (DS).[32] Both were shown in the literature to produce crystal-like conformers for a set of drug-like compounds.[14−16,18] However, for the present training data set, MacroModel had fewer compounds with no bioactive-like conformations (12) in comparison with DS (19) and more compounds without any nonbioactive conformers (24 vs 19 for DS). Consequently, we decided to base our algorithm on the conformations generated by MacroModel.

Prefiltration Stage. By carefully examining our data as well as those available in the literature, we found that for some ligands, bioactive conformers could not be obtained using reasonable efforts with current conformational search methods.[14] While improving conformational search algorithms goes beyond the scope of the current work, we searched for a simple prefiltration criterion that would filter out those compounds for which the probability of identifying bioactive-like conformations is low. Since the ability to identify bioactive-like conformations is likely to decrease with ligand flexibility, we focused on two relevant 2D descriptors, namely, the number of RotBond already invoked in the past for a similar purpose[16,21] and the newly developed M-PRoB; (see Methods and Figure 2).

Table 2a presents the number of ligands in the training set with at least one bioactive-like conformation as a function of M-PRoB. As ligand flexibility increases, the probability of finding bioactive-like conformations is expected to decrease. This is clearly evidenced in the case of conformations generated by DS but less so in those generated by MacroModel, where an oscillatory behavior is observed beyond M-PRoB > 8. Whether this behavior reflects a real phenomena (e.g., since more conformations are generated for more flexible ligands, the probability that one of them would be bioactive-like increases) or results from the smaller number of lignds with M-PRoB > 8 in our data set is still an open question that merits further work. Table 2b presents the number of ligands having more than 50% of their conformations as nonbioactive as a function of M-PRoB. Here, the expected positive correlation between these two values is observed with both DS- and MacroModel-generated

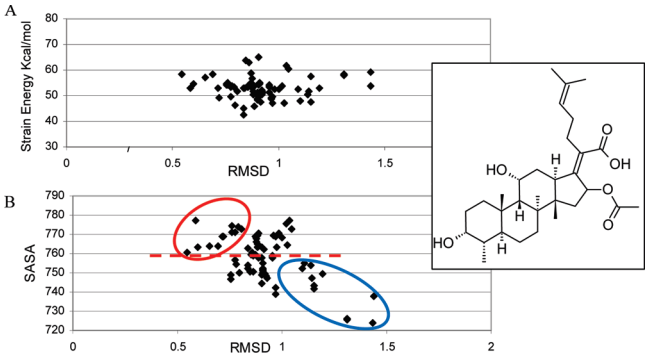

Figure 4. (a) Strain energy as a function of rmsd from the crystal structure for the set of 3D conformations of fucidic acid as obtained by MacroModel. (b) Solvent accessible surface area (SASA) as a function of rmsd from the crystal structure for the set of 3D conformations of fucidic acid as obtained by MacroModel. The 2D structure of fucidic acid is shown in the insert.

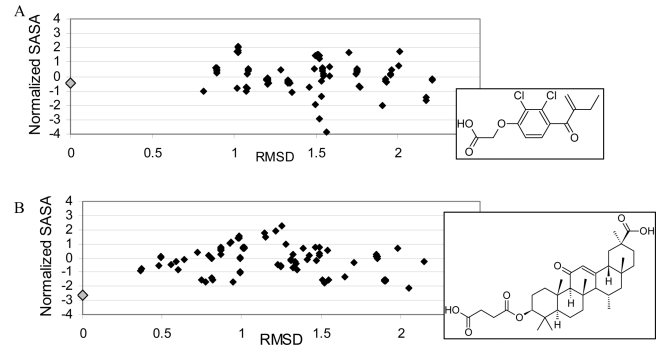

Figure 5. (a) Normalized solvent accessible surface area (SASA) as a function of rmsd from the crystal structure for the set of 3D conformations of etacrynic acid as obtained by MacroModel. (b) Normalized solvent accessible surface area (SASA) as a function of rmsd from the crystal structure for the set of 3D conformations of cerbenoxolone as obtained by MacroModel. The 2D structures of etacrynic acid (top) and cerbenoxolone (bottom) are shown in the inserts.

conformations. By looking at these data, we have decided to set a prefiltration threshold criterion of M-PRoB ≥ 8 for MacroModel-generated conformations. Indeed, this threshold filters out 10 compounds and only 3 of which have bioactive-like conformations. Moreover, those 10 compounds have a total of 4 393 nonbioactive conformations for an average of 439 nonbioactive conformations per compound. This should be compared with the average of 126 nonbioactive conformations per compound across the entire training set.

Due to the high correlation observed in the present data set between M-PRoB and RotBond (data not shown), it is instructive to look at similar data as a function of RotBond (see Table 2c and d). In accord with previous reports,[14,16,21] the prefiltration threshold is set at RotBond ≥9, which results in removal of 13 compounds, 3 of which have bioactive-like conformations. Thus, for this training set, M-PRoB was less restrictive than RotBond and left us with more compounds.

Filtration Stage. Having identified M-PRoB ≥ 8 as the prefiltration criterion, we now turn to the actual filtration process. Initially, we sought to identify a single 3D descriptor that would be able to single out bioactive-like conformations for all or most of the compounds in our training set. Consider as an example Figure 4a and 4b that present, respectively, the correlation between the strain energy and the solvent accessible surface area (SASA) and the rmsd with respect to the crystal structure for the set of MacroModel generated conformations of fusidic acid. While strain energy is clearly not correlated with rmsd, selecting a cutoff filter of SASA > 760 Å^2 will retain the best rmsd conformations with values less than 1 Å (red ellipse), while filtering out the conformations with the highest rmsd (blue ellipse). However, SASA does not perform as well in other cases, e.g., etacrynic acid (Figure 5a) and carbenoxolone (Figure 5b). In both cases, neither the raw SASA values (data not shown) nor the normalized ones are correlated with rmsd (only normalized values will be considered from this point to allow for a comparison between 3D different descriptors). All other 3D descriptors considered in this work showed a similar behavior. In particular, in our hands and for the present data set, the energetic criterion was unable to consistently distinguish between bioactive-like and nonbioactive conformations, although such a criterion was invoked in the past.[16]

Having failed to identify a single 3D descriptor as a global filter, we reasoned that partitioning of the training set into multiple groups followed by filtration of each group with a unique 3D descriptor might lead to a good filtration model. In order to avoid a potential overfitting problem, we decided to limit the initial division of the training set to two groups only. Consider for example Figure 6a and b. The bioactive-like conformations of cyclothiazide are primarily found in bin-6 of the normalized radius of gyration descriptor, whereas those of flufenamic acid are primarily in bins-4 and -5. Thus, selecting either bin-6 or bins-4 and -5 will inevitably lead to a loss of bioactive-like conformations for one of these compounds. However, as the molecular weight (MW) of cyclothiazide (390) is much higher than that of flufenamic acid (280), we can use the molecular weight as a splitting descriptor followed by picking bin-6 for high molecular weight compounds (e.g., MW > 350) and bins-4 and -5 for low molecular weight compounds (e.g., MW < 350). Obviously, selecting two filtration criteria to distinguish between two compounds will lead to gross overfitting. However, we hoped that we could find a simple 2D and 3D (termed 2D-3D-3D) combination that would enrich bioactive-like conformations in the entire training set.

Model Development. With this in mind, we developed a set of filtration models consisting of a general workflow (Figure 3) and a unique combination of normalized 2D-3D-3D descriptors selected through an exhaustive search in the large descriptor combination space. Model generation proceeded by testing each 2D-3D-3D combination for its ability to meet the predefined success criteria discussed above. These success criteria were expressed as combinations of threshold values and are presented in Table 3 together with the number of output models satisfying each combination. Due to the large number of possible 2D-3D-3D combinations, it is not surprising that, generally, each thresholds combination was satisfied by multiple models.

A total of 45 2D-3D-3D models (each defined by the identity and the value of the 2D and 3D descriptors) met the most favorable set of success criteria (row 5 of Table 3) when applied to the 47 training set compounds. These models were further filtered by the goodness-of-hit list (eqs 3−6), passing only models with a goodness-of-hit list >0.17 for both groups and an average value >0.19. Only 12 models satisfied these criteria and were characterized by the following 2D splitting descriptors: numbers of bonds, aromatic

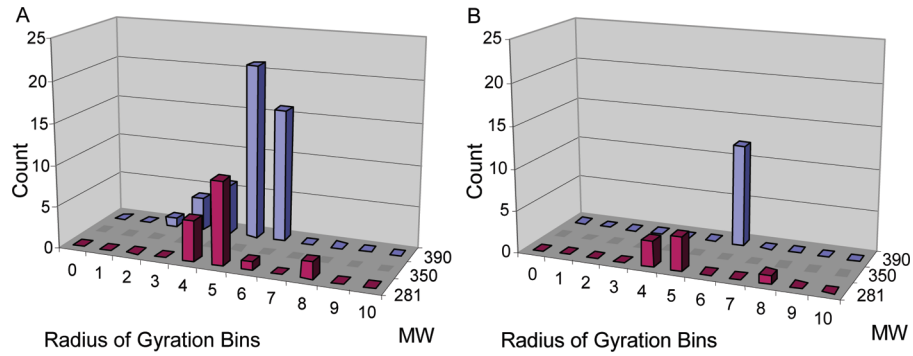

Figure 6. (a) All MacroModel-generated conformations of flufenamic acid (red) and cyclothiazide (blue) organized according to the 2D molecular weight descriptor and binned according to the normalized radius of gyration. (b) Bioactive-like conformations of flufenamic acid (red) and cyclothiazide (blue) organized according to the 2D molecular weight descriptor and binned according to the normalized radius of gyration.

Table 3. Combinations of Success Criteria Expressed as a Set of Threshold Values Tested by Our Procedure and Their Output[a]

	average bioactive-like enrichment cutoff	average nonbioactive impoverishment cutoff	no. molecules with sufficient bioactive-like conformations	no. models
1	1.00	0.40	37	148
2	1.20	0.40	37	0
3	1.15	0.40	37	8
4	1.15	0.40	35	61
5	1.20	0.40	35	45
6	1.00	0.25	35	122
7	1.00	0.25	35	57

[a] Averaged bioactive-like enrichment and nonbioactive impoverishment values over all training set compounds were calculated according to eq 2. Column 4 gives the tolerance cutoff, namely, the number of compounds for which the success criteria must be met. The last column gives the number of output models satisfying all criteria within the pre-defined tolerance.

rings, H-bond donors, Wiener index, flexibility index (PHI), and ALogP.

Seven of the models used ALogP as the 2D splitting descriptor, and all resulted in the subsequent, 3D descriptor-based selection of almost all populated bins for the final subset. Consequently, these models were expected to have a limited "filtration power", which was indeed confirmed by testing their performance of the independent test set (data not shown). This can be due to the fact that ALogP does not describe structural complexity; therefore, a molecule with 6 atoms can have the same LogP as a molecule with 6 000 atoms. Molecular complexity is better described by the other 2D splitting descriptors that were finally used. Thus, LogP, while being one of the most important descriptors in drug discovery, should not be used for a crude filtration.

The remaining five models are presented in Table 4. Of these, model 4 uses the number of aromatic rings as the 2D splitting descriptor, setting the split point at 1.5 (namely, splitting is based on whether the number of aromatic rings is 0−1 or larger than 1). However, we expect the usefulness of this criterion to be limited during lead optimization since the number of aromatic ring systems is not expected to change significantly during analog synthesis, and consequently, we did not use this model.

Models 1 and 2 are similar differing only in their 2D splitting descriptor (Model 1: number of bonds; Model 2: Wiener index; see below). The Wiener index describes the molecular complexity as the sum of the chemical bonds between all pairs of heavy atoms in the molecule. Interestingly, both resulting model groups used SASA as the 3D descriptor albeit with different bins. For Wiener index values <520, the model selects all SASA bins except 3 and 4, whereas for Wiener index values ≥520, the model only selects bins-5 and -6. Thus bins-5 and -6 are selected for both groups of the Wiener index, suggesting that there is a higher probability to find bioactive-like conformations with average to moderately high SASA. This observation is consistent with previous findings.[26] A similar analysis extended to the 38 models meeting the most stringent success criteria (out of the 45 models described above excluding ALogP-based models) has reached similar conclusions, suggesting that for about 75% of the compounds, the bioactive-like conformations are expected to have medium to high SASA values. Nevertheless, for about a quarter of the compounds a different criterion is needed.

An illustrative example for model 3 is presented in Figure 7. This model utilizes PHI as the 2D splitting descriptor (pink in Figure 7). Compounds with PHI ≥ 3.65 are assigned to the right (blue) part and binned according to the PMI-mag descriptor. Only conformations residing in bins 0−4 and 6−10 are collected. Those compounds having PHI < 3.65 are assigned to the left (yellow) part and binned according to SASA. Only conformations residing in bins-5 and -6 are collected.

Model Performance on the Independent Test Set. *Prefiltration.* The performances of the filtering models were evaluated using an independent test set consisting of 24 compounds. In the prefiltration stage, the most critical issue is the sensitivity, namely, the ability to retain all compounds that have bioactive-like conformations. Five compounds were prefiltered by the M-PRoB ≥ 8 criteria. Indeed, MacroModel did not produce bioactive-like conformations for four of these compounds and for the fifth, only a single bioactive-like conformation was identified among its 690 conformations (which also included 615 nonbioactive conformations). The prefiltration sensitivity success rate is, therefore, 80%. The total number of conformations removed by prefiltration was 2 887, out of which 2 476 (86%) were classified as nonbioactive. The prefiltration procedure did not filter two compounds with no bioactive-like conformations that had a total number of 844 conformations. Similar results were obtained by using the RotBond ≥ 9 criterion as a prefilter. Thus, RotBond ≥ 9 is a viable alternative to M-PRoB.

While we envision the main role of prefiltration as a tool for removing from subsequent processing compounds for which the probability of identifying bioactive-like conforma-

Table 4. Five Models with the Best Goodness-of-Hit List Value from Among the 45 Models Satisfying the Success Criteria Listed in Raw 5 of Table 3[a]

model	2D splitting descriptor	splitting value	group	3D descriptor	normalized bins selected	compounds in selected bins	goodness-of-hit list	bioactive-like enrichment	nonbioactive impoverishment
1	no. of bonds	<18.5	1	Jurs-SASA	0−2, 5−10	9	0.22	1.12	0.00
		>18.5	2	Jurs-SASA	5−6	28	0.17	1.23	0.35
2	Wiener	<520	1	Jurs-SASA	0−2, 5−10	9	0.22	1.12	0.00
		≥520	2	Jurs-SASA	5−6	28	0.17	1.12	0.35
3	PHI	<3.65	1	PMI-mag	0−4, 6−10	8	0.22	2.42	0.00
		≥3.65	2	Jurs-SASA	5−6	29	0.21	1.13	0.34
4	aromatic rings	<1.5	1	Jurs-SASA	5−9	24	0.22	1.19	0.24
		>1.5	2	PMI-mag	5−7	13	0.13	1.42	0.20
5	HB-donors	<3.5	1	Jurs-SASA	0−2, 5−10	26	0.17	1.22	0.19
		>3.5	2	shadow-nu	5	11	0.23	1.14	0.30

[a] All models used the M-PRoB ≥ 8 as the pre-filtering criterion. The 2D splitting descriptors and their corresponding splitting values into two groups are listed in the "2D splitting descriptor" and "splitting value" columns. The "3D descriptor" column lists the descriptors used for the binning of each of the two groups. For each group, the selected bins are given in the "normalized bins selected" column, and the resulting number of compounds is found in the "compounds in selected bins" column. The "goodness-of-hit list" values for each group were calculated according to eqs 3−6, and the bioactive-like conformations enrichment and non-bioactive conformations impoverishment were calculated according to eq 2.

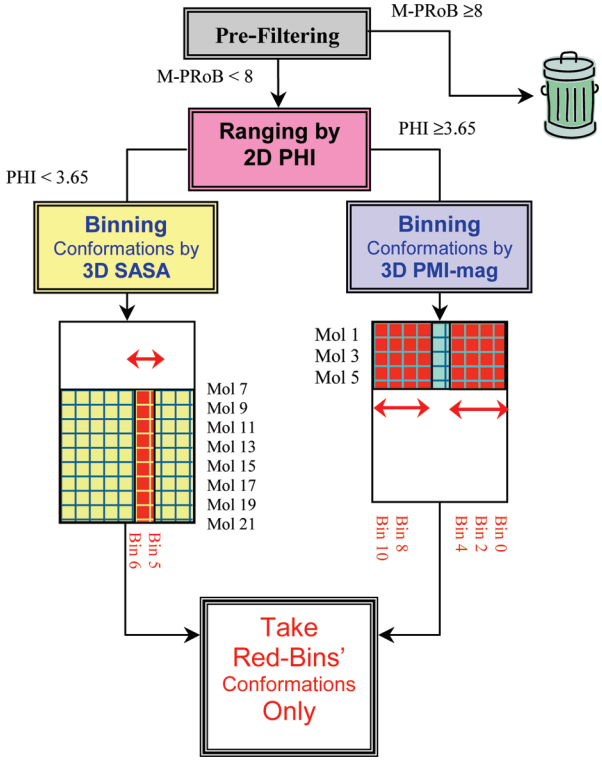

Figure 7. Schematic representation of the filtration workflow employed by model 3. The prefiltration step uses the M-PRoB ≥ 8 criterion. The remaining compounds are split into two groups based on the PHI value. Compounds with PHI ≥ 3.65 are assigned to the right (blue) group and binned according to the PMI-mag descriptor. Only conformations residing in bins 0−4 and 6−10 are collected. Those compounds having PHI < 3.5 are assigned to the left (yellow) group and binned according to SASA. Only conformations residing in bins-5 and -6 are collected.

tions in small, it is still instructive to look at its performance when combined with subsequent filtration steps (see below). Thus, the entire filtration model reduced the total number of conformations by 61% and the proportion of nonbioactive conformations from 46 to 20%. At the same time, the proportion of bioactive-like conformations was increased from 4 to 7%.

Performance of Model 1 (Number of Bonds-SASA-SASA) and of Model 2 (Wiener Index-SASA-SASA) on the Test Set. Using the number of bonds and the Wiener index as splitting descriptors resulted in an identical division of the training set compounds into two groups and, consequently, led to a practical identity between models 1 and 2. Thus, the results of these two models on the independent test set are discussed together. Both models are presented in Table 4, and their performances on the test set are summarized in Tables 5 (model 1) and 6 (model 2). The total number of conformations, the number of conformations classified as nonbioactive, and the number of conformations classified as bioactive-like, which were removed by this second filtration step, were 1 666 (36%), 377 (39%), and 107 (35%), respectively. Overall, the 3D-based filtration considerably reduced the number of conformations while affording a slight enrichment of the bioactive-like ones and an impoverishment of nonbioactive conformations. Of the remaining 19 compounds that were not prefiltered by the M-PRoB ≥ 8 criterion, 5 presented only bioactive-like conformations (pyrimethamine, diethylstilbestrol, efavirenz, dexamethasone, acetazolamide; see Tables 1 and 3 of the Supporting Information; these 5 compounds are all small and/or rigid with a "narrow" conformational space), and consequently, the enrichment cutoff value of 1.2 could not be mathematically met. Nevertheless, we consider these cases as successes since the filtration procedure may have resulted in the removal of all of their conformations including the bioactive-like ones. As 10 additional compounds (acetylcholine, mitoguazone, ibuprofen, flurbiprofen, adenosine, ribostamycin, calcipotriol, melagatran, kanamycin, acarbose), met the success criteria (retention of a sufficiently large number of bioactive-like conformation and bioactive-like enrichment ≥1.2 or nonbioactive impoverishment ≤0.4), the success rate of the second filtration step is 15/19 or 79%, and that of the entire model (including prefiltration) is 19/24 or 79% (although this latter number is not rigorously defined as the success criteria of the prefiltration, and the second filtration steps are different).

A detailed analysis of the data presented in Tables 5 and 6 reveals encouraging trends. In most cases (except sulfasalazine and erythromycin), we were able to retain a sufficiently large number of bioactive-like conformations, which would allow for ligand-based approaches in lead

Table 5. Performance of Model 1 on the External Test Set of 19 Compounds Following Pre-Filtration[a]

	before 3D filtration			after 3D filtration					
	all	bioactive-like	nonbioactive	all	bioactive-like	nonbioactive	% removed	bioactive-like enrichment	nonbioactive impoverishment
ACH	115	2	0	53	2	0	54	2.2	NA
MGB	18	7	0	11	5	0	39	1.2	NA
AZM	23	23	0	17	17	0	26	1.0	NA
IBP	120	64	0	85	55	0	29	1.2	NA
CP6	4	4	0	3	3	0	25	1.0	NA
FLP	69	53	0	14	10	0	80	0.9	NA
DES	3	3	0	1	1	0	67	1.0	NA
ADN	97	15	0	71	15	0	27	1.4	NA
EFZ	2	2	0	2	2	0	0	1.0	NA
SAS	363	15	33	161	3	33	56	0.5	2.3
DEX	47	47	0	32	32	0	32	1.0	NA
MC9	471	28	29	314	28	4	33	1.5	0.2
MEL	140	0	0	96	0	0	31	NA	NA
RIO	363	17	81	198	17	36	45	1.8	0.8
KAN	323	2	53	236	2	47	27	1.4	1.2
SPP	925	8	244	590	3	160	36	0.6	1.0
KTN	704	0	172	523	0	155	26	NA	1.2
ACR	539	5	360	339	5	160	37	1.6	0.7
ERY	250	15	1	164	3	1	34	0.3	1.5

[a] For the full name of the compounds see Table 1 in the Supporting Information. Each compound (column 1) is characterized by the total number of conformations (column 2), the number of conformations classified as bioactive-like (column 3), and the number of conformations classified as nonbioactive (column 4) prior to the 3D-based filtration. Information pertaining to post filtration is reported in columns 5−7. Column 8 reports the percentage of removed compounds, while columns 9 and 10, respectively, report the enrichment of bioactive-like conformations and the impoverishment of nonbioactive conformations as calculated by eq 2. The upper part of the table (above the bold line) refers to the first group emerging from the 2D splitting, whereas the lower part refers to the second group. NA indicates compounds for which bioactive-like enrichment and/or non-bioactive impoverishment calculations have no meaning.

Table 6. Performance of Model 2 on the External Test Set of 19 Compounds Following Pre-Filtration[a]

	before 3D filtration			after 3D filtration					
	all	bioactive-like	nonbioactive	all	bioactive-like	nonbioactive	% removed	bioactive-like enrichment	nonbioactive impoverishment
ACH	115	2	0	53	2	0	54	2.2	NA
AZM	23	23	0	17	17	0	26	1.0	NA
MGB	18	7	0	11	5	0	39	1.2	NA
IBP	120	64	0	85	55	0	29	1.2	NA
CP6	4	4	0	3	3	0	25	1.0	NA
FLP	69	53	0	14	10	0	80	0.9	NA
AND	97	15	0	71	15	0	27	1.4	NA
EFZ	2	2	0	2	2	0	0	1.0	NA
DES	3	3	0	1	1	0	67	1.0	NA
DEX	47	47	0	32	32	0	32	1.0	NA
SAS	363	15	33	161	3	33	56	0.5	2.3
RIO	363	17	81	198	17	36	45	1.8	0.8
MC9	471	28	29	314	28	4	33	1.5	0.2
MEL	140	0	0	96	0	0	31	NA	NA
KAN	323	2	53	236	2	47	27	1.4	1.2
SPP	925	8	244	590	3	160	36	0.6	1.0
KTN	704	0	172	523	0	155	26	NA	1.2
ACR	539	5	360	339	5	160	37	1.6	0.7
ERY	250	15	1	164	3	1	34	0.3	1.5

[a] See legend of Table 5 for more details.

optimization. At the same time, both the total number of conformations and the number of conformations classified as nonbioactive (when present in the input conformational ensemble) were significantly reduced. However, there is clearly room for improvement.

Performance of Model 3 (PHI-PMI$_{magnitude}$-SASA) on the Test Set. Model 3 is presented in Table 4, and its performances on the test set are summarized in Table 7. The total number of conformations, the number of conformations classified as nonbioactive, and the number of conformations classified as bioactive-like, which were removed by this second filtration step, were 1 719 (38%), 377 (39%), and 115 (37%), respectively. A total of 14 compounds (74%) met the success criteria discussed above (acetazolamide, adenosine, pyrimethamine, efavirenz, acetylcholine, ibuprofen, dexamethasone, diethylstilbestrol, mitoguazone, calcipotriol, melagatran, ribostamycin, kanamycin, acarbose). The overall success rate of the entire model (including prefiltration) is 18/24 or 75%.

Performance of Model 5 (Number of H-bond Donors-SASA-SASA) on the Test Set. Model 5 is presented in Table 4 and its performances on the test set are summarized in Table 8.

Table 7. Performance of Model 3 on the External Test Set of 19 Compounds Following Pre-Filtration[a]

	before 3D filtration			after 3D filtration					
	all	bioactive-like	nonbioactive	all	bioactive-like	nonbioactive	% removed	bioactive-like enrichment	nonbioactive impoverishment
AZM	23	23	0	23	23	0	0	1.0	NA
AND	97	15	0	64	13	0	34	1.3	NA
CP6	4	4	0	1	1	0	75	1.0	NA
EFZ	2	2	0	1	1	0	50	1.0	NA
FLP	69	53	0	18	2	0	74	0.1	NA
ACH	115	2	0	3	2	0	97	38.3	NA
IBP	120	64	0	84	54	0	30	1.2	NA
DEX	47	47	0	32	32	0	32	1.0	NA
DES	3	3	0	1	1	0	67	1.0	NA
MGB	18	7	0	9	5	0	50	1.4	NA
SAS	363	15	33	161	3	33	56	0.5	2.3
SPP	925	8	244	590	3	160	36	0.6	1.0
MC9	471	28	29	314	28	4	33	1.5	0.2
KTN	704	0	172	523	0	155	26	NA	1.2
MEL	140	0	0	96	0	0	31	NA	NA
RIO	363	17	81	198	17	36	45	1.8	0.8
KAN	323	2	53	236	2	47	27	1.4	1.2
ACR	539	5	360	339	5	160	37	1.6	0.7
ERY	250	15	1	164	3	1	34	0.3	1.5

[a] See legend of Table 5 for more details.

Table 8. Performance of Model 5 on the External Test Set of 19 Compounds Following Pre-Filtration[a]

	before 3D filtration			after 3D filtration					
	all	bioactive-like	nonbioactive	all	bioactive-like	nonbioactive	% removed	bioactive-like enrichment	nonbioactive impoverishment
ACH	115	2	0	53	2	0	54	2.2	NA
EFZ	2	2	0	2	2	0	0	1.0	NA
FLP	69	53	0	26	10	0	62	0.5	NA
IBP	120	64	0	85	55	0	29	1.2	NA
CP6	4	4	0	3	3	0	25	1.0	NA
AZM	23	23	0	17	17	0	26	1.0	NA
DES	3	3	0	1	1	0	67	1.0	NA
KTN	704	0	172	530	0	162	25	NA	1.3
DEX	47	47	0	33	33	0	30	1.0	NA
MC9	471	28	29	325	28	4	31	1.5	0.2
SPP	925	8	244	638	5	173	31	0.9	1.0
AND	97	15	0	37	7	0	62	1.2	NA
ERY	250	15	1	96	5	0	62	0.9	0.0
SAS	363	15	33	149	1	13	59	0.2	1.0
MEL	140	0	0	46	0	0	67	NA	NA
MGB	18	7	0	10	5	0	44	1.3	NA
RIO	363	17	81	144	4	19	60	0.6	0.6
KAN	323	2	53	113	0	14	65	0.0	0.8
ACR	539	5	360	148	0	94	73	0.0	1.0

[a] See legend of Table 5 for more details.

The total number of conformations, the number of conformations classified as nonbioactive, and the number of conformations classified as bioactive-like, which were removed by this second filtration step, were 2 120 (46%), 494 (51%), and 132 (43%), respectively. In contrast with the above models, only 11 compounds passed the success criteria (58%; acetylcholine, efavirenz, flurbiprofen, ibuprofen, pyrimethamine, acetazolamide, dexamethasone, calcipotriol, adenosine, melagatran, mitoguazone). The overall success rate of the entire model (including prefiltration) is 15/24 or 62%.

This model utilized the number of H-bond donors as the 2D splitting descriptor, setting the split point at 3.5. For compounds having H-bond donors <3.5 (upper part of Table 8), subsequent filtering using the SASA descriptor led to satisfactory results, similar to those obtained with models 1−3. However, for compounds having H-bond donors >3.5, the situation is different. Subsequent filtering using the shadow-nu 3D descriptor focused on a single bin and filtered out 61% of the conformations but also 59% of the bioactive-like ones, leaving two out of eight compounds (kanamycin and acarbose; 25%) with no bioactive-like conformations at all and a third (sulfasalazine) with only a single bioactive-like conformation out of a total of 15 in the input ensemble. Based on these results, filtration criteria which result in the focus on a single bin seem too restrictive and lead to poorer performance when applied to external test sets.

Significance of 2D Splitting. Finally, we tested whether using a 2D splitting descriptor had a beneficial effect on model performance. We focused our attention on the very similar models 1 and 2 (see Table 3) since both used the same 3D descriptor in the two groups resulting from the 2D split, thereby, allowing for an easy comparison with corresponding unsplit

models. Two unsplit models were considered, corresponding to the bins selected by the SASA descriptor for split group 1 (bins 0, 1, 2, and 5−10; see Table 4) and for split group 2 (bins 5, 6; see Table 4). Using bins-5 and -6 only to filter the entire unsplit data set led to the loss of the two bioactive-like conformations of acetylcholine and to the loss of one bioactive-like conformation for ibuprofen. On the other hand, using bins 0, 1, 2, and 5−10 to filter the entire unsplit data set introduced an additional 144 conformations, 46 of which are nonbioactive. These results, therefore, demonstrate the benefit obtained by the 2D-based splitting.

CONCLUSIONS AND FUTURE DIRECTIONS

The current work was motivated by the observation that ligand-based lead discovery and optimization campaigns tend to spend excessive resources analyzing irrelevant ligand conformations. We, therefore, developed a new filtration procedure to eliminate such conformations while maintaining a sufficiently large number of bioactive-like conformations. Our models were developed from a diverse training set of 47 drug compounds and tested on an independent test set of 24 drug compounds, thereby, providing a nonbiased estimate for model performance.

The results presented in this work are encouraging. First, prefiltration successfully removed five compounds, with unsatisfactory conformational ensembles from the 24 compounds in the test set. Subsequent filtration steps showed, by themselves, good results, in particular those obtained with models 1 and 2. Thus, for the remaining 19 compounds, an average of 36% of the total conformations and 39% of conformations classified as nonbioactive were removed. Furthermore, for 15 out of the 17 compounds with at least one bioactive-like conformation, a sufficiently large number of such conformations, was retained enabling the reliability of ligand-based modeling. The two misses of this approach are sulfasalazine and erythromycin. However, even for these compounds, 33−56% of the total conformations were filtered out while 3 good bioactive-like conformations, out of 15, were retained.

Following prefiltration, the majority of the conformations (72%), of the remaining compounds, are neither bioactive-like nor nonbioactive ($2.5 > \text{rmsd} > 1$ Å). The total number of these conformations was reduced by the filtration procedure by 36%, although their proportion in the filtered ensemble of the test set compounds did not change. Still, as the analysis of this mass of conformations requires significant attention and resources during lead discovery and optimization, we consider this reduction as significant.

Since there are many methods by which conformational ensembles could be generated and each could, in principle, be used in conjunction with many force fields, we did not try to exhaustively evaluate all methods. Rather, we limited our attempts to those conformational search methods already shown in the literature to provide reasonable conformational ensembles[14−16,18] and to force fields which could be applied to the large diversity of potential drug-like compounds without the need to tediously derive new parameters. We reason that as the quality of the input ensemble improves, so will the results of the current procedure (although model rederivation might be required).

Clearly, more work is justified to further refine and test the procedure described herein. In particular, more exhaustive assessment of the conformational ensembles obtained by other conformational search methods for larger data sets and evaluation of models derived from additional 2D and 3D descriptors and their combinations would be beneficial. Moreover, such filters, once developed, should be integrated into conformational search and docking algorithms in order to bias their results toward the bioactive region of the conformational space. Work along these lines will be undertaken in our laboratory in the near future.

ACKNOWLEDGMENT

We are thankful to Dr. Katalin Nadassy from Accelrys Inc. for developing the protocols for the calculation of the new descriptor mentioned in the paper and for all of her assistance and support. We also thank Dr. Roberto Olender and Dr. Boaz Inbal from EPIX Pharmaceutical for Perl scripting guidance (R.O.) and the many fruitful discussions (R.O., B.I.).

Supporting Information Available: A list of ligands used in the present work, a list of training set ligands and their structures, a list of test set ligands and their structures, a list of proteins from the SuperDrug database, and a description of the descriptors used in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44* (7), 1035–42.
(2) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–63.
(3) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862–5.
(4) Guido, R. V.; Oliva, G.; Andricopulo, A. D. Virtual screening and its integration with modern drug design technologies. *Curr. Med. Chem.* **2008**, *15* (1), 37–46.
(5) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7* (20), 1047–55.
(6) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–73.
(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **200**, *28*, 235–42.
(8) Prathipati, P.; Dixit, A.; Saxena, A. K. Computer-Aided Drug Design: Integration of Structure-Based and Ligand-Based Approaches in Drug Design. *Curr. Comput.-Aided Drug Des.* **2007**, *3* (2), 133–148.
(9) Guner, O. F. History and evolution of the pharmacophore concept in computer-aided drug design. *Curr. Top. Med. Chem.* **2002**, *2* (12), 1321–32.
(10) Leach, A. R. A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B. , Eds.; Wiley-VCH: New York, 1991; Vol. 2, 1−55.
(11) Schwab, C. H. Conformational Analysis and Searching. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed. Wiley-VCH: Weinheim, Germany, 2003; Vol. 1, 262301.
(12) Howard, A. E.; Kollman, P. A. An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.* **1988**, *31* (9), 1669–75.
(13) Bostrom, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15* (12), 1137–52.
(14) Good, A. C.; Cheney, D. L. Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. *J. Mol. Graph. Model.* **2003**, *22* (1), 23–30.
(15) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47* (10), 2499–510.
(16) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. In.f Model.* **2005**, *45* (2), 422–30.

(17) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, *47* (3), 1067–86.
(18) Chen, I. J.; Foloppe, N. Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2008**, *48* (9), 1773–91.
(19) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.* **2003**, *21* (5), 449–62.
(20) Borodina, Y. V.; Bolton, E.; Fontaine, F.; Bryant, S. H. Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space. *J. Chem. Inf. Model.* **2007**, *47* (4), 1428–37.
(21) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J. Comput. Chem.* **2006**, *27* (16), 1962–9.
(22) Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tuffery, P. Frog: a FRee Online druG 3D conformation generator. *Nucleic Acids Res.* **2007**, *35* (Web Server issue), W568–72.
(23) Stockwell, G. R.; Thornton, J. M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **2006**, *356* (4), 928–44.
(24) Gunther, S.; Senger, C.; Michalsky, E.; Goede, A.; Preissner, R. Representation of target-bound drugs by computed conformers: implications for conformational libraries. *BMC Bioinformatics* **2006**, *7*, 293.
(25) Diller, D. J.; Merz, K. M., Jr. Can we separate active from inactive conformations? *J. Comput.-Aided Mol. Des.* **2002**, *16* (2), 105–12.
(26) Michalsky, E.; Dunkel, M.; Goede, A.; Preissner, R. SuperLigands - a database of ligand structures derived from the Protein Data Bank. *BMC Bioinformatics* **2005**, *6*, 122.
(27) *Discovery Studio*, version 2.1; Accelrys Inc., San Diego, CA; www.accelrys.com.
(28) Chang, G.; Wayne C. Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.
(29) *Maestro 7.5*; Schrödinger Inc., Portland, OR; www.schrodinger.com.
(30) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C. Macro-Model - an Integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics. *J. Comput. Chem.* **1990**, *11* (4), 440–467.
(31) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting conformational variation. *J. Comput. Chem.* **1995**, *16*, 171–187.
(32) Triballeau, N.; Bertrand, H.-O.; Acher, F. , Are You Sure You Have a Good Model? In *Pharmacophores and Pharmacophore Searches*; Langer, T.; Hoffmann, R. D. , Eds. Wiley-VCH: Weinheim, Germany, 2006; Vol. 32, 338−340.
(33) Vedani, A.; Dobler, M. 5D-QSAR: the key for simulating induced fit. *J. Med. Chem.* **2002**, *45*, 2139–49.
(34) Wu, G.; Robertson, D. H.,; Brooks, C. L., III; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24* (13), 1549–62.

CI900163T