# Analysis of a High-Throughput Screening Data Set Using Potency-Scaled Molecular Similarity Algorithms

Ingo Vogt and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Received December 4, 2006

Molecular similarity methods for ligand-based virtual screening (VS) generally do not take compound potency as a variable or search parameter into account. We have incorporated a logarithmic potency scaling function into two conceptually distinct VS algorithms to account for relative compound potency during search calculations. A high-throughput screening (HTS) data set containing cathepsin B inhibitors was analyzed to evaluate the effects of potency scaling. Sets of template compounds were randomly selected from the HTS data and used to search for hits having varying potency levels in the presence or absence of potency scaling. Enrichment of potent compounds in small subsets of the HTS data set was observed as a consequence of potency scaling. In part, observed enrichments could be rationalized as a result of recentering chemical reference space on a subspace populated by potent compounds. Our findings suggest that VS calculations using multiple reference compounds can be directed toward the preferential detection of potent database hits by scaling compound contributions according to potency differences.

## INTRODUCTION

For the in silico screening of compound databases, different types of methods have been adopted, including techniques for quantitative structure−activity relationship (QSAR) analysis,[1] pharmacophore methods,[2] and approaches that are based on the more global concept of molecular similarity.[3] QSAR methods correlate structural features and properties of molecules with their activity.[1] The paradigm of QSAR analysis is to suggest small structural modifications that significantly improve the biological activity of test compounds. Therefore, QSAR analysis requires the presence of discontinuous structure−activity relationships, but exploring such SARs can also give rise to significant errors.[4] Although multidimensional QSAR models have been adapted for compound database mining,[5] the field of ligand-based virtual screening (VS) is presently dominated by molecular similarity methods that are conceptually based on the *similar property principle*[3] and employ a holistic molecular view.[6] In addition, 3D database search techniques focusing on pharmacophore models are also widely used.[2,7] Given their whole-molecule perspective, molecular similarity methods do not make any assumption about pharmacophores or parts of molecules that render them biologically active[6] and can thus be applied when little or no SAR information is available. Similarity methods require the presence of continuous SARs where departures from the structures of active compounds cause gradual changes in biological activity, consistent with the *similar property principle*. In contrast to QSAR analysis, neither pharmacophore nor similarity methods usually take differences in compound potency into account. The qualitative manner in which SARs are explored causes a limitation of similarity methods: newly identified hits are generally much less potent than the reference molecules because one deliberately departs from optimized structural motifs.[6,8] Given this situation, we have investigated the inclusion of compound potency as a search parameter in VS calculations to tune them toward the detection of potent database hits. We first incorporated a potency scaling function into the *Dynamic Mapping of Consensus Positions* (DMC) algorithm[9] and observed that this modification led to a relative enrichment of potent compounds among correctly identified hits for three activity classes.[10] On the basis of these findings, we have further improved the potency-scaled DMC algorithm (POT−DMC) and added potency scaling to a recently developed and completely different method, a distance function that navigates high-dimensional descriptor spaces, *Distance in Activity-Centered Chemical Space* (DACCS),[11] and created POT−DACCS. These algorithms were applied to mine a publicly available high-throughput screening (HTS) data set with a relatively narrow potency distribution of hits. This data set presented a fairly challenging test case, as further discussed below. In POT−DMC and POT−DACCS, the same potency scaling function was incorporated into distinct algorithms, and both methods were found to enrich potent compounds in HTS subsets when compared to nonscaled calculations, which suggests a more general applicability of potency scaling in VS analysis.

## METHODS

**Algorithms.** We selected two different VS algorithms for potency scaling, DMC and DACCS. Outlines of DMC and DACCS are shown in Figures 1 and 2, respectively. DMC operates in descriptor spaces of increasing dimensionality that are simplified through median-based binary descriptor transformation,[12] which converts property descriptors with continuous value ranges into a binary format based on the

* Corresponding author tel: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci6005432 CCC: $37.00 
Published on Web 02/15/2007

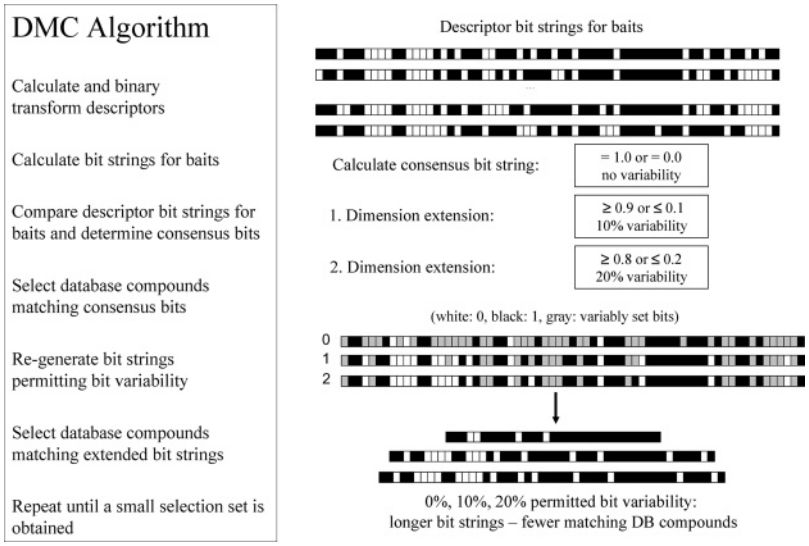


**Figure 1.** DMC algorithm. A summary of the algorithm is shown, as described in the text. "DB" stands for database, and "baits" refers to known active reference compounds.

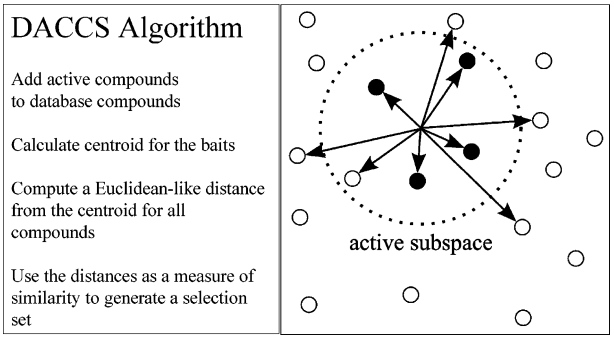


**Figure 2.** DACCS algorithm. Black dots represent a subset of reference molecules used to calculate the centroid of an active subspace (dotted circle) and open dots database compounds located close to or within the subspace. Arrows indicate Euclidian-like distances of compounds from the centroid.

statistical medians of their value distributions in a screening database. A compound is assigned a value of 1 for a specific descriptor if its actual value is larger or equal to the median or a 0 if it is smaller. This binary model makes it possible to generate descriptor bit strings for the mapping of compounds. A consensus position is defined by a descriptor vector composed of bit positions having an identical setting for all reference compounds. Dimension extension is achieved by defining consensus positions that no longer require identical descriptor settings for all templates. Dimension extension levels 1, 2, and 3 permit 10%, 20%, and 30% variability in descriptor bit settings, respectively. Mapping to activity-dependent consensus positions in descriptor spaces of increasing dimensionality removes most database compounds from these consensus positions and retains only the most similar ones. The DMC algorithm has produced significant hit and recovery rates in a number of test cases.[9]

DACCS is designed to operate in high-dimensional but unmodified descriptor spaces and takes contributions of many arbitrarily chosen descriptors into account. Through a scaling procedure, DACCS centers chemical reference space on a subspace populated by a set of reference molecules and approximates this subspace as an orthogonal system by calculation of Euclidian-like distances from the center of the subspace to all compounds in a database. Thus, DACCS calculations generate a distance-based ranking of database compounds from the centroid of an active subspace, in contrast to DMC, which produces variably sized compound selection sets, depending on the number of database compounds mapping to consensus positions. Thus, DMC and DACCS are algorithmically distinct.

**Potency Scaling.** In order to scale the contributions of reference molecules to define consensus positions or active subspaces relative to their potencies, a logarithmic scaling function was applied:

$$SF = \ln(pot_{min}) - \ln(pot_{act}) + 1$$

Here, $pot_{min}$ is the lowest potency occurring in the set of reference molecules and $pot_{act}$ the potency of an individual reference compound. The addition of 1 ensures that the compound having the lowest potency in the set is assigned a scaling factor (SF) of 1. Compounds having higher potency are assigned larger scaling factors. For example, if the lowest potency within the reference set is 100 nM, the SF of a compound having 10 nM potency would be $(4.6 - 2.3 + 1) = 3.3$ and the SF of a compound having 1 nM potency would be $(4.6 - 0 + 1) = 5.6$, and so on. Thus, every reference compound is assigned an individual scaling factor.

In addition to logarithmic scaling, we have previously analyzed different scaling schemes. Direct scaling by potency values is often prohibitive because the search calculations are dominated by the most potent or a few highly potent compounds, which presents a problem for the analysis of continuous potency distributions. An alternative to logarithmic scaling would be the application of linear scaling functions where continuous scaling factors are assigned to represent a potency distribution. However, these scaling factors put more weight on compounds having midrange potency than those having either high or low potency. Therefore, such linear scaling procedures do not effectively direct DMC or DACCS calculations toward the recognition of potent database hits. For the DMC and DACCS algorithms, a logarithmic scaling function is usually preferred because it emphasizes contributions from highly potent compounds and takes balanced contributions from com-

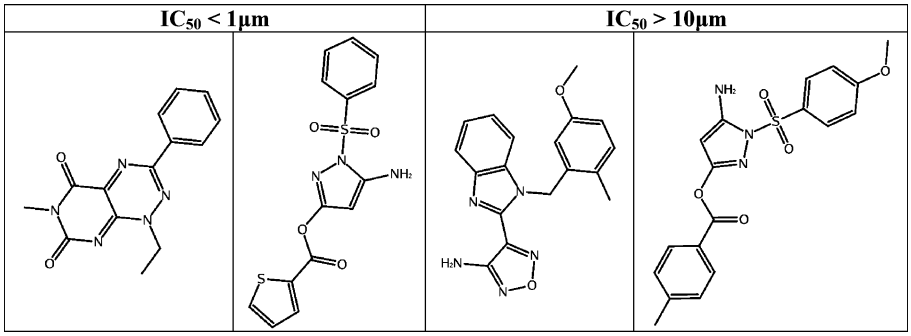


**Figure 3.** Examples of cathepsin B inhibitor screening hits. Shown are examples of hits having different potencies.

pounds covering the entire potency range into account. Only the compound having lowest potency is not scaled.

In POT−DMC, the logarithmic scaling factors become effective at the level of calculating consensus bits. If a descriptor bit is set on for a reference compound, it is multiplied by its individual scaling factor. Then, scaled bit values are summed at each bit position and normalized by the sum of all scaling factors (to obtain values between 0 and 1). This simply means that bit settings of highly potent compounds are counted several times more than those of weakly potent ones when calculating consensus bits. Thus, more potent compounds have a higher weight in defining consensus positions. During dimension extension, bit settings of highly potent compounds become statistically more decisive for the acceptance of additional descriptors.

In the original implementation of POT−DMC,[10] we followed the dimension extension scheme of DMC, as described above. However, we observed substantial changes in compound numbers between single dimension extension steps, which made it difficult to monitor potential potency enrichments because all database compounds were deselected after only one to three steps in our test calculations.[10] In order to alleviate these problems, we introduce a "smooth" dimension extension function for POT−DMC, which makes it possible to control the reduction in compound numbers during mapping to consensus positions and continuously monitor potency distributions. The underlying idea is that the *variability of descriptor bit positions can only adopt discrete values depending on the number of available reference compounds*. For DMC, the increase in bit variability $v$ per dimension extension step is set to the *smallest possible change in bit distribution*:

$$v = 1/\mathrm{ref}$$

with ref being the number of reference molecules. For POT−DMC, bit variability per step is set to

$$v = 1/\sum_{i}^{\mathrm{ref}} \mathrm{SF}_i$$

with $\mathrm{SF}_i$ being the potency scaling factor of reference molecule $i$. Thus, for 10 available reference compounds, the permitted bit variability per step would be 10% in DMC calculations, but for 100 available reference compounds, it would only be 1%. In POT−DMC calculations, potency scaling becomes effective during dimension extension, not when defining the initial consensus position.

In DACCS calculations, potency scaling operates at a different level. Given a set of known active reference compounds, the distance in scaled chemical space ($d_{\mathrm{DACCS}}$) from the center of the "active subspace" is calculated for database compounds as follows:

$$d_{\mathrm{DACCS}} = \sqrt{\sum_{i}^{d}\left(\frac{x_i - \overline{\mathrm{act}_i}}{\mathrm{stdev}_i}\right)^2}$$

Here $x_i$ is the value of one of the $d$ descriptors for a compound $x$, $\overline{\mathrm{act}_i}$ the mean value of descriptor $i$ for a set of active template compounds, and $\mathrm{stdev}_i$ the standard deviation of the descriptor values for the templates. If this standard deviation is zero, then the standard deviation of the entire compound population is used instead. If both values are zero, the descriptor is omitted from the calculation (which would apply to a descriptor having the same value for all active and database compounds).

In POT−DACCS, potency scaling becomes effective when the means and standard deviations are calculated for centering the active subspace. Under scaling conditions, descriptor values of potent compounds with large scaling factors determine the means and standard deviations more than weakly potent compounds with small scaling factors. Thus, the descriptor statistics change through the weight put on the values of more potent compounds (essentially, these values are counted several times more than the ones of weakly potent molecules). As a consequence, the active subspace is shifted toward the positions of potent compounds.

Typically, compound potencies are considered in QSAR analysis in order to derive functions that relate molecular structure and properties to differences in compound activity and ultimately predict increasingly potent analogs.[1] By contrast, similarity methods are in general qualitative in nature and do not take differences in potency into account.[6] To our knowledge, POT−DMC and POT−DACCS are the first similarity-based methods that include relative compound potency as a search parameter.

**HTS Data and Calculations.** The HTS data set was generated at the Penn Center for Molecular Discovery at the University of Pennsylvania[13] using an assay for inhibitors of cathepsin B, a cysteine protease in lysosomes, and made public through PubChem.[14] It consists of 63 332 compounds including 40 hits with $IC_{50}$ values ranging from 46 nM to 46 $\mu$M. Examples of hits are shown in Figure 3. The structural resemblance of active compounds was evaluated on the basis of average pairwise Tanimoto coefficients (Tc)[15] calculated with a fingerprint consisting of the publicly

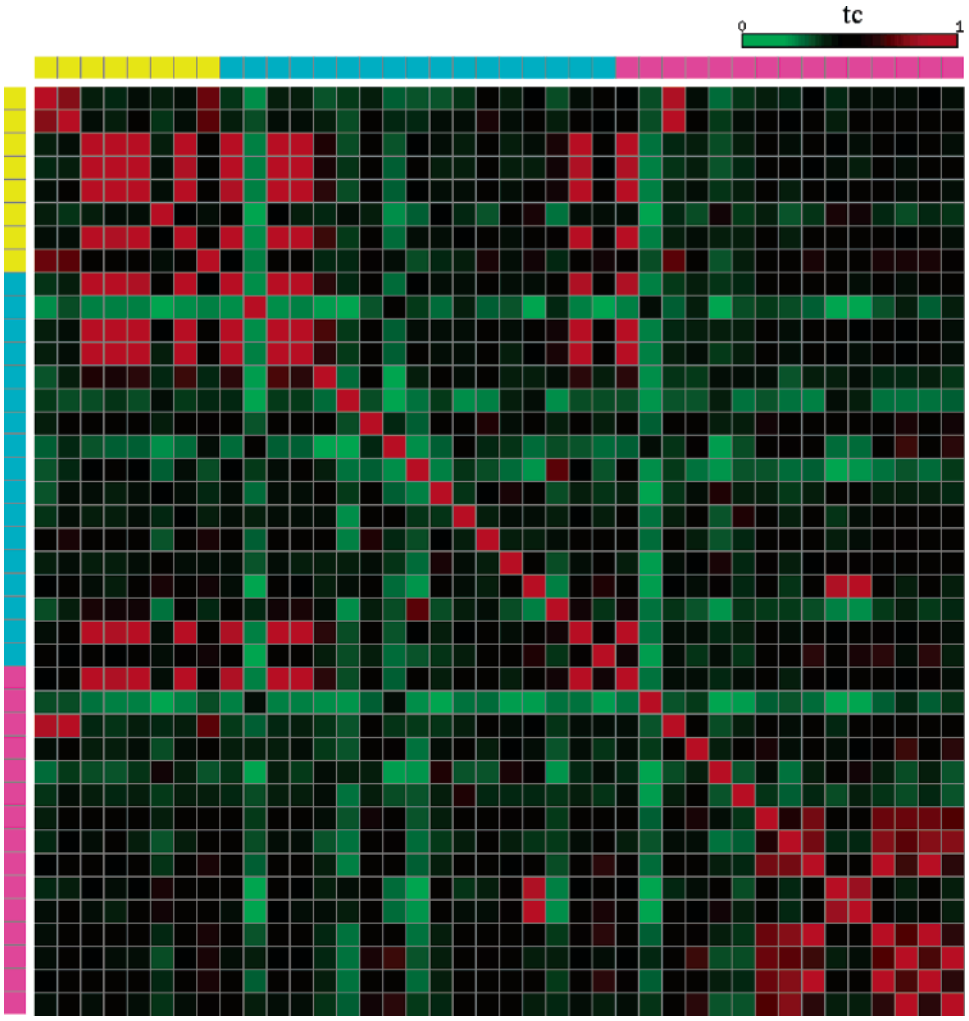


**Figure 4.** Tc matrix of hits. MACCS Tc values were calculated for pairwise comparison of all active compounds. For Tc values, continuous color-coding is used with increasing similarity from green to red (black indicates Tc = 0.5). Color bars mark the matrix positions of compounds grouped by different potency levels: <1 $\mu$M (yellow), between 1 and 10 $\mu$M (cyan), >10 $\mu$M (magenta). From top to bottom and left to right, compounds are listed in the order of decreasing potency. The Tc matrix was generated using Matrix2png.[24]

available set of 166 MACCS structural keys.[16] The similarity of active and inactive molecules was assessed using single template search calculations with MACCS for each known active against the remainder of the database.

For VS trials, 30 sets of 10 active compounds each were randomly selected as reference compounds (so that 30 active compounds remained as hits in the database). The only preselection requirement was that the reference compounds had to cover the experimentally observed potency range (in order to ensure realistic potency scaling). Each set of reference compounds was individually used as input for scaled and nonscaled DACCS and DMC calculations. For search calculations, a set of 155 1D/2D molecular descriptors was calculated with MOE.[17] For DMC and POT−DMC calculations, descriptor medians were calculated for the HTS data set.

## RESULTS AND DISCUSSION

**Characterization of HTS Data.** We first analyzed the distribution of hits in the HTS set and the similarity of active and inactive compounds. Three compounds were active below 100 nM and five additional ones below 1 $\mu$M; 17 hits had potency in the range between 1 and 10 $\mu$M, and the remaining 15 hits were active above a 10 $\mu$M concentration. Thus, the potency distribution of active molecules was continuous but relatively narrow, with most compounds being in the low micromolar range. The mean and median potencies were 13.2 and 6.9 $\mu$M, respectively. Highly potent (e.g., low nanomolar) compounds are rarely found in primary screening data sets. However, despite its moderate average potency, the cathepsin B set contained a total of eight sub-micromolar hits including three highly active compounds. Figure 4 reports the results of pairwise MACCS Tc comparisons of potent compounds. In the Tc matrix, off-diagonal red patterns indicate significant similarity between hits. The figure reveals that similarity between hits was limited to relatively few (of 780 possible) compound pairs. Many comparisons yielded Tc values of 0.5 or smaller, which are indicative of dissimilar compounds in MACCS key calculations. Similarity was greatest among seven potent (top-left red cluster) and seven weakly active (lower right) compounds. The two most active compounds were similar to each other, but also to a weakly active molecule, and compounds in the top-left cluster were also found to display distinct similarity to two weakly active ones. Therefore, compound potency within this set did not clearly correlate with compound similarity (as also illustrated by the example structures in Figure 3), and in addition, active compounds were not confined to one or two analog series

**Table 1.** Similarity between Hits and Inactive Compounds in the Cathepsin B Data Set[a]

| $IC_{50}$ [nm] | 1.00 | (1.00, 0.95) | (0.95, 0.90) | (0.90, 0.85) | (0.85, 0.80) |
|---|---|---|---|---|---|
| 46.1 | 0 | 0 | 0 | 1 | 7 |
| 71.0 | 0 | 0 | 0 | 0 | 11 |
| 72.2 | 0 | 0 | 0 | 1 | 14 |
| 246.6 | 0 | 5 | 5 | 14 | 43 |
| 434.6 | 0 | 2 | 7 | 8 | 16 |
| 691.9 | 0 | 5 | 8 | 19 | 105 |
| 844.5 | 0 | 0 | 0 | 0 | 3 |
| 923.5 | 0 | 5 | 7 | 16 | 92 |
| 1185.1 | 1 | 3 | 22 | 47 | 64 |
| 1260.8 | 0 | 2 | 6 | 8 | 14 |
| 1639.5 | 0 | 0 | 0 | 0 | 0 |
| 1749.1 | 0 | 5 | 7 | 12 | 75 |
| 1990.0 | 0 | 6 | 4 | 17 | 71 |
| 2087.4 | 1 | 0 | 5 | 0 | 1 |
| 2120.0 | 0 | 0 | 0 | 0 | 0 |
| 2247.2 | 1 | 0 | 6 | 53 | 217 |
| 3170.9 | 0 | 0 | 6 | 16 | 62 |
| 4169.7 | 0 | 0 | 0 | 1 | 1 |
| 6355.6 | 0 | 0 | 1 | 3 | 8 |
| 6714.5 | 0 | 1 | 10 | 16 | 29 |
| 7113.8 | 0 | 0 | 0 | 2 | 16 |
| 8563.1 | 0 | 0 | 3 | 19 | 102 |
| 8926.9 | 0 | 0 | 1 | 17 | 38 |
| 9388.7 | 0 | 0 | 0 | 1 | 13 |
| 9562.6 | 2 | 5 | 10 | 37 | 132 |
| 11 457.8 | 0 | 0 | 0 | 1 | 4 |
| 12 265.1 | 2 | 6 | 12 | 56 | 144 |
| 12 947.1 | 0 | 0 | 1 | 1 | 0 |
| 14 196.4 | 0 | 0 | 0 | 3 | 23 |
| 18 349.1 | 0 | 1 | 3 | 12 | 34 |
| 19 686.4 | 0 | 0 | 0 | 2 | 3 |
| 28 132.2 | 0 | 0 | 2 | 5 | 48 |
| 33 855.4 | 2 | 1 | 1 | 1 | 5 |
| 37 190.0 | 0 | 0 | 0 | 2 | 20 |
| 38 470.6 | 1 | 0 | 0 | 1 | 4 |
| 39 986.8 | 0 | 0 | 1 | 1 | 19 |
| 44 181.1 | 0 | 0 | 1 | 2 | 27 |
| 44 577.0 | 0 | 1 | 3 | 56 | 237 |
| 44 782.6 | 0 | 0 | 0 | 3 | 29 |
| 46 161.2 | 0 | 0 | 3 | 44 | 160 |
| **sum** | **10** | **48** | **135** | **498** | **1891** |

[a] For each hit, listed by potency, the number of inactive compounds matching Tc intervals between 0.80 and 1.00 is reported.

because intraset structural similarity would be higher in that case. Table 1 reports the results of MACCS similarity search calculations within the screening set using each hit individually as a query. A considerable number of inactive data set compounds were found to be structurally similar or very similar to hits, including nearly identical compounds (MACCS Tc = 1.00). At a MACCS Tc level of 0.90 or greater, nearly 200 matches of hits to inactive compounds were detected. At a MACCS Tc threshold value of ~0.85, the structural resemblance of compounds was clearly visible, and nearly 700 matches were detected at this level. These findings nicely illustrate that Tc values alone (irrespective of the fingerprints used) are not a reliable indicator of similar activity[18] and that very similar structures can either be active or inactive.[19] The latter point is of course well-known in medicinal chemistry and analog design[19] and presents an intrinsic limitation of methods that evaluate similarity from a whole-molecule perspective.[20,21]

The relatively narrow and continuous potency distribution in the cathepsin B HTS set, the limited number of available hits, and the structural diversity among active compounds made this data set a challenging test case for our potency-scaled similarity analysis. In general, narrow potency distributions and diverse active compounds make potency-oriented molecular similarity calculations difficult. By contrast, the presence of discontinuous potency distributions over several orders of magnitude and structural homogeneous subsets of highly and weakly potent compounds make it easy to direct search calculations toward the recognition of the most potent database hits through potency scaling procedures.

**Experimentally Oriented Benchmarking.** The structural diversity among hits reflected in Figure 4 and the similarity of active and inactive molecules indicated in Table 1 made this screening set a rather challenging test case for similarity methods. However, it provided a much more realistic benchmarking situation than one could typically generate when sets of known active compounds are added to a background database for test calculations, for several reasons: (1) compound activity classes obtained from the literature or other sources usually consist of optimized and highly active molecules that are often not representative of hits identified in screening campaigns; (2) like HTS, VS can only aim at identifying hits, not leads or drugs; (3) background database compounds must be considered decoys because their potential activities are unknown, whereas HTS data sets contain confirmed negatives; (4) moderately sized HTS screening sets are often biased or heterogeneous in nature (e.g., through opportunistic compound sourcing), which can complicate screening calculations. Thus, even if we consider the likely presence of experimental errors or noise in HTS data sets, analyzing such data sets more closely mimics practical VS conditions than a typical benchmark scenario, and efforts to make HTS data sets publicly available through PubChem[14] or other initiatives[22,23] provide a significant opportunity for the VS field.

Given the features of the cathepsin B HTS set described above, we asked two major questions in our study: (a) can we recover active molecules from this experimental set through DMC and DACSS calculations, and if so, (b) is it possible to enrich selections with potent database compounds using POT−DMC and POT−DACCS? In other words, can similarity calculations be directed toward the recognition of potent hits when mining experimental screening data?

**DMC and DACCS Calculations.** We first analyzed individual nonscaled search calculations using different sets of reference compounds in order to determine whether active molecules could be retrieved from the HTS set, irrespective of their potency levels. Results obtained for DMC and DACCS are reported in Tables 2 and 3, respectively. DMC calculations consistently found active compounds at the final dimension extension levels (i.e., before no compounds in the data set mapped to further extended consensus bit stings). In a few cases, the selection sets contained a large number of inactive compounds together with two or three hits, for example, in trials 13 and 21 in Table 2. For practical VS applications, such calculations would not be suitable because too many database compounds would need to be evaluated. However, most calculations produced reasonably sized selection sets, and in nine trials, one or two inhibitors were recovered together with only approximately 10 or even fewer inactive compounds.

The majority of DACCS calculations reported in Table 3 also detected inhibitors. Five trials failed to produce active

**Table 2.** Hits in DMC and POT−DMC Compound Selection Sets[a]

| trial | DMC | POT−DMC |
|---|---|---|
| 1 | 5/435 | 1/221 |
| 2 | 2/196 | 3/39<br>2/12 |
| 3 | 2/4<br>1/2 | 3/24<br>2/9 |
| 4 | 1/41 | 1/34 |
| 5 | 2/97<br>1/15 | 2/36<br>1/3 |
| 6 | 2/24 | 4/9<br>2/4 |
| 7 | 1/11 | 1/2 |
| 8 | 2/663 | 2/92<br>1/19 |
| 9 | 1/13 | 2/36<br>1/4 |
| 10 | 8/749 | 2/285 |
| 11 | 1/223 | 1/160 |
| 12 | 2/269 | 1/70 |
| 13 | 2/5320 | 2/1669 |
| 14 | 1/93 | 1/7 |
| 15 | 1/23<br>1/4 | 2/29<br>1/4 |
| 16 | 1/466 | 1/9 |
| 17 | 3/189<br>2/4 | 3/82<br>2/9<br>1/2 |
| 18 | 2/98 | 2/493 |
| 19 | 1/7 | 2/37<br>1/2 |
| 20 | 1/49 | 2/14 |
| 21 | 3/780 | 2/152 |
| 22 | 2/64 | 3/8 |
| 23 | 1/62 | 1/3 |
| 24 | 3/36<br>2/4 | 2/4<br>2/2 |
| 25 | 1/170 | 2/5 |
| 26 | 1/20 | 3/55<br>1/2 |
| 27 | 2/80 | 2/60<br>1/4 |
| 28 | 1/61<br>1/8 | 2/36<br>1/2 |
| 29 | 1/60 | 3/23<br>1/12 |
| 30 | 2/49 | 4/47<br>2/12 |

[a] Results of 30 individual search calculations with DMC and POT−DMC are summarized. For each calculation, hits identified at the final or second to last dimension extension level are reported. For example, "2/24" means that the compound selection set contained two cathepsin B inhibitors and 22 inactive compounds.

**Table 3.** Hits in DACCS and POT-DACCS Compound Selection Sets[a]

| | DACCS | | POT−DACCS | |
|---|---|---|---|---|
| trial | S50 | S100 | S50 | S100 |
| 1 | 0 | 1 | 3 | 4 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 4 | 5 | 5 | 5 |
| 4 | 0 | 0 | 2 | 3 |
| 5 | 5 | 6 | 5 | 5 |
| 6 | 1 | 1 | 1 | 4 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 6 | 6 | 5 | 5 |
| 9 | 0 | 0 | 1 | 1 |
| 10 | 2 | 4 | 3 | 4 |
| 11 | 4 | 4 | 6 | 6 |
| 12 | 0 | 1 | 0 | 0 |
| 13 | 5 | 6 | 6 | 6 |
| 14 | 0 | 2 | 1 | 1 |
| 15 | 0 | 0 | 1 | 2 |
| 16 | 0 | 0 | 1 | 1 |
| 17 | 2 | 3 | 3 | 3 |
| 18 | 3 | 4 | 4 | 5 |
| 19 | 4 | 5 | 5 | 5 |
| 20 | 1 | 1 | 0 | 0 |
| 21 | 0 | 1 | 1 | 2 |
| 22 | 0 | 1 | 1 | 1 |
| 23 | 6 | 6 | 5 | 5 |
| 24 | 0 | 0 | 1 | 1 |
| 25 | 0 | 0 | 0 | 0 |
| 26 | 1 | 1 | 3 | 3 |
| 27 | 5 | 5 | 5 | 5 |
| 28 | 0 | 2 | 1 | 2 |
| 29 | 0 | 1 | 0 | 0 |
| 30 | 0 | 0 | 0 | 0 |

[a] For individual DACCS and POT-DACCS calculations, hits contained in selection sets of 50 (S50) or 100 (S100) database compounds are reported.

molecules among the top-scoring 100 compounds, but 14 calculations revealed between one and six hits among the top 50 compounds with, on average, 3.5 inhibitors per trial. The remaining calculations produced between one and six active molecules among the top 100 compounds.

In general, the results of DMC and DACCS calculations were much influenced by the composition of the reference molecule sets, which is also observed in VS benchmark calculations[6] and enforces the need to explore multiple reference sets in order to obtain a more general picture about the performance of a method. Under the experimental benchmark conditions investigated here, both DMC and DACCS produced lower hit rates than observed in the initial evaluations of these methodologies, as we had anticipated. When tested on five different activity classes, DMC produced hit rates between 9% and 74%[9] and DACCS between 15% and 60%,[11] although it should be noted that the calculation of hit rates might become artificial when compound selection sets become very small, as often observed with DMC.[9] It should also be considered that, different from benchmarking, hit and recovery rates are not a primary measure of success for VS applications because compound recovery cannot be determined in practical screens and because the ability to identify novel active molecules in small VS selection sets is much more relevant than actual hit rates, which represent a more suitable measure for large-scale screening campaigns. Since both DMC and DACCS calculations consistently retrieved active molecules from the HTS data sets, we were able to explore our key question of whether or not increasingly potent inhibitors would be detected under the conditions of potency scaling.

**DMC versus POT−DMC.** Applying the new dimension extension function, the size of selection sets could be much better controlled than in the original implementation of POT−DMC.[10] The comparison in Table 2 shows that DMC and POT−DMC calculations detected an overall comparable number of hits. However, POT−DMC recognized fewer inactive molecules in 17 of 30 calculations, which considerably reduced the size of the selection sets. In a number of cases, the reduction in the number of inactive molecules was quite dramatic, for example, in trials 8, 21, or 25 (Table 2). Thus, potency scaling increased the specificity of the calculations. Next, we analyzed the potency distribution of the identified inhibitors. POT−DMC detected on average 1.3 hits with potency $<1\ \mu$M in selection sets of fewer than

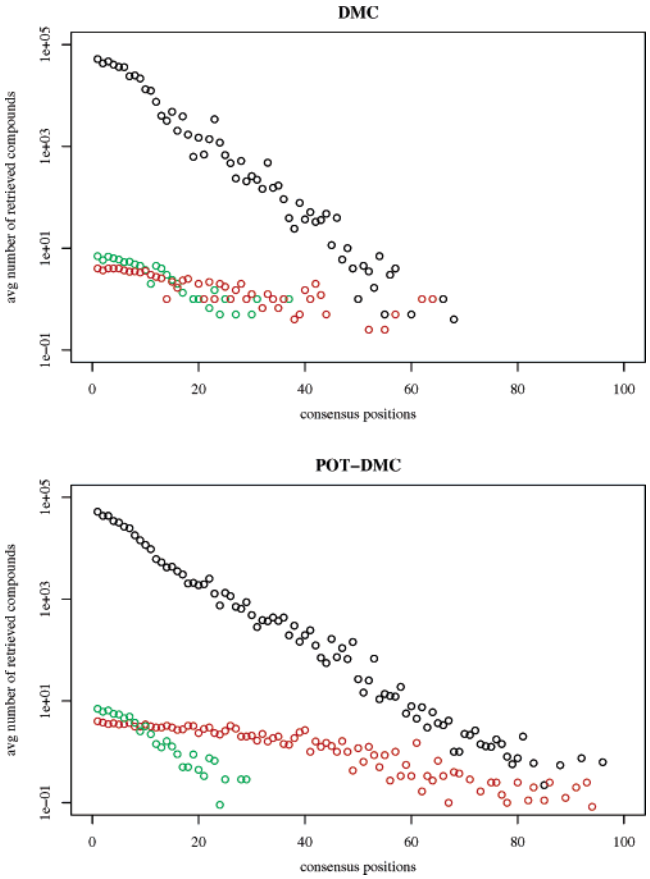


**Figure 5.** Potency distribution of hits at increasing dimension extension levels. Shown is a comparison of compounds matching search strings of increasing length in DMC and POT−DMC calculations. During dimension extension, the number of descriptor consensus bits increases ("consensus positions" reports the number of accepted descriptor bits). Black circles represent database compounds, green circles hits with >10 $\mu$M potency, and red circles hits with <1 $\mu$M potency.

100 compounds, whereas DMC detected on average 0.6 of these hits. The POT−DMC retrieval rate was considerable because only four potential hits with <1 $\mu$M potency were available in the screening set. POT−DMC also detected approximately twice as many hits as DMC with a potency of up to 10 $\mu$M but did not detect hits with >10 $\mu$M potency. Thus, there was a consistent trend of POT−DMC calculations to enrich selection sets with more potent compounds than DMC.

Underlying effects could be studied by analyzing the mapping pathways of hits with different potencies to consensus bit strings during dimension extension, as reported in Figure 5. For DMC, there were only little differences in the mapping characteristics of active compounds, although some potent compounds reached higher dimension extension levels than the weakly potent molecules. By contrast, for POT−DMC, significant changes were observed. Here, weak hits were eliminated earlier during the search than in DMC calculations, but the most potent hits reached much higher dimension levels than in nonscaled calculations, which resulted in a distinct separation of more and less potent compounds during POT−DMC calculations. In both cases, search strings consisting of a small number of consensus bits produced comparable results because, at the beginning of the calculations, only low bit variability is permitted, which suppresses the scaling effect. The observation that potent hits were retained under increasingly stringent search conditions during POT−DMC calculations can be explained when considering the fact that POT−DMC generates consensus bit settings that selectively favor highly potent molecules.

**DACCS versus POT-DACCS.** Next, we analyzed the effects of POT−DACCS, which is algorithmically distinct from POT−DMC because it produces a distance-based ranking of database compounds, with increasing distance in chemical space being a measure of decreasing similarity. As shown in Table 3, POT−DACCS calculations identified more hits than DACCS in approximately half of the trials. In all of the trials that produced hits, POT−DACCS achieved higher average potency than DACCS. For example, among the top scoring 100 data set compounds, POT−DACCS recognized on average one of the four most potent hits available in the database, whereas DACCS recognized on average only ~0.3. Similar to the observations made for POT−DMC, POT−DACCS displayed a tendency to deselect weakly active hits relative to DACCS and generally did not rank them among the top scoring 100 compounds.

In contrast to POT−DMC, compound ranking also needs to be taken into account when evaluating POT−DACCS. Therefore, we compared the potency versus rank distributions of correctly identified hits in POT−DACCS and DACCS calculations and found that POT−DACCS produced an enrichment of potent compounds relative to DACCS in 21 of 27 search calculations that accumulated hits in the top 1% of the data set. Representative examples are shown in Figure 6 where it can be seen that POT−DACCS calculations concentrate hits in the lower-left quarter of the graphs, the area corresponding to highly ranked potent hits. On the basis of systematic rank versus potency comparisons, two effects were found to be responsible for the enrichment. POT−DACCS ranked potent hits more highly than DACSS and assigned lower ranks to weakly potent hits, which resulted in an enrichment of most potent hits at higher rank positions, as illustrated in Figure 6. Over all of the search calculations, the average rank position of the most potent hits within the top scoring 100 compounds was 25 for DACCS and 19 for POT−DACCS, and for hits with potency between 1 and 10 $\mu$M, the average rank was 23 for DACCS and 20 for POT−DACCS. For hits with potency >10 $\mu$M, an average rank of 34 was determined for DACCS, whereas POT−DACCS did not select weakly potent hits among the top 100 compounds. These findings confirmed that POT−DACCS assigned lower ranks to weakly active hits than DACCS and higher ranks to increasingly potent hits.

Given the findings discussed above, we attempted to illustrate the effects of potency scaling in an intuitive manner. Since DACCS operates in unmodified descriptor spaces, we have systematically searched for combinations of three property descriptors that would produce a notable separation of more and less potent hits in a three-dimensional representation. An example is shown in Figure 7. Here, mapping a random sample of 5000 inactive compounds from the HTS data set into a space constituted by three intuitive descriptors (logP, molecular weight, and approximate van der Waals volume) produced a visible separation of more and less potent inhibitors. Of course, this descriptor combination would not be sufficient for VS because it did not separate active molecules from the "cloud" of inactive compounds. However, despite its limited resolution, this reference space made

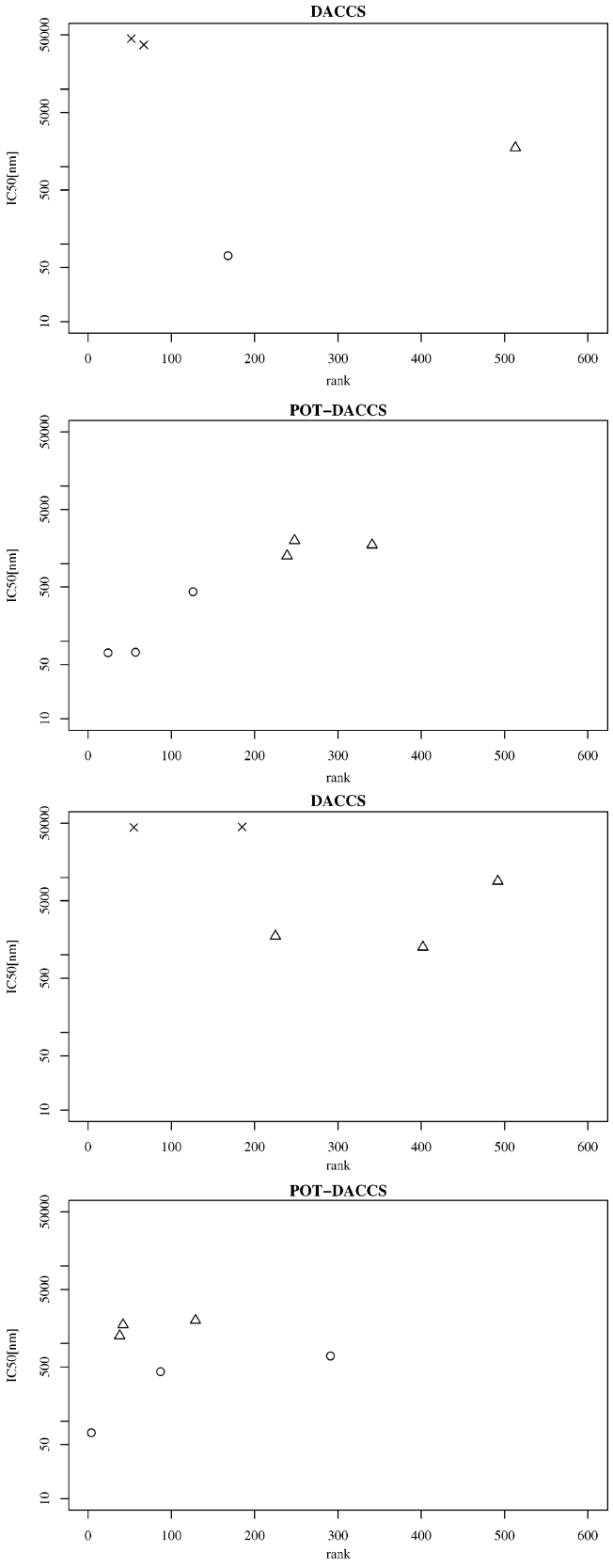


**Figure 6.** Potency distribution of hits in DACCS and POT−DACCS trials. For two sets of template compounds, the potency distribution of hits in corresponding DACCS and POT−DACSS calculations is compared. Plotted is the potency of hits relative to their rank in the priority list for ∼1% of the database. Active compounds are assigned to potency levels as in Figure 4: <1 $\mu$M (dots), between 1 and 10 $\mu$M (triangles), >10 $\mu$M (crosses). Enrichment of potent hits is indicated by the accumulation of active compounds in the lower left section of the graph.

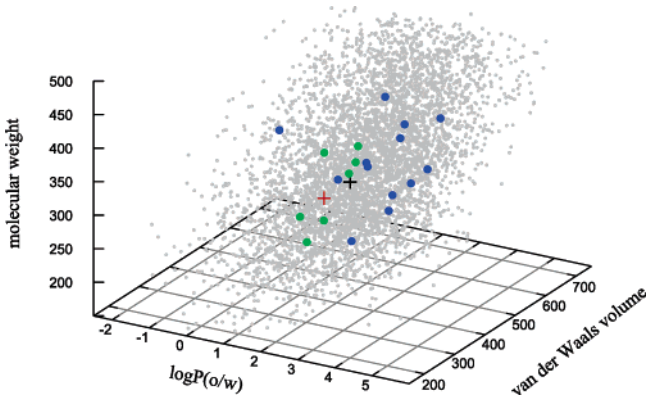


**Figure 7.** Compound distributions and centroids. A sample of the cathepsin B HTS data set was projected into a three-dimensional descriptor space. Hits with potency <1 $\mu$M are represented as green dots, hits with potency >10 $\mu$M as blue dots, and inactive compounds as small gray dots. Average centroid positions for all DACCS (black cross) and POT−DACCS calculations (red) are mapped.

it possible to visualize the effects of potency scaling on the location of the centroid position. Figure 7 reveals that potency scaling shifted the DACCS centroid position away from weakly potent hits into a region predominantly occupied by molecules having higher potency, which rationalizes why POT−DACCS calculations detected more potent inhibitors. Therefore, we can at least in part attribute the effects of potency scaling to recentering of the activity-dependent subspaces. It is reasonable to assume that similar changes in centroid positions also occur in high-dimensional space representations when potency scaling is applied.

**Perspective.** We have shown that two conceptually different similarity-based algorithms could be successfully modified to take compound potency as a calculation parameter into account and preferentially select increasingly potent hits. In both cases, a logarithmic scaling function was used that assigned scaling factors to active compounds according to their potency. Molecular similarity methods are typically qualitative in nature, and POT−DMC and POT−DACCS are the first similarity methods that explicitly consider relative compound potency during database searching. For DMC and POT−DMC, a practical advancement has been the introduction of a smooth dimension extension function that alleviates previous problems associated with dramatic changes in compound numbers during the first one to three dimension extension steps. Since DMC and DACCS are different methods, it is conceivable that potency scaling could also be applied to similarity search tools or compound classification methods other than our two algorithms.

## CONCLUSIONS

We have described an approach to incorporate compound potency as a parameter in similarity calculations. For the evaluation of these calculations, we regarded the availability of an experimental screening data set as very important because its analysis mimics VS conditions much better than many conventional benchmark settings. The cathepsin B HTS data set we studied contained a limited number of hits with a narrow but continuous potency distribution. Limited similarity among active compounds and the presence of inactive molecules that were very similar to hits made this

a challenging test case. Nevertheless, DMC and DACCS calculations using alternative sets of reference molecules consistently retrieved active molecules, in part, in very small selection sets. Both POT−DMC and POT−DACCS calculations favored the recognition of potent hits and displayed a tendency to deselect weakly potent ones, while retrieving similar or larger numbers of hits from the HTS data set compared to nonscaled calculations. Thus, we conclude that potency scaling, as implemented in POT−DMC and POT−DACCS, can successfully extrapolate from the features of potent reference molecules and direct VS calculations toward the recognition of potent hits.

## REFERENCES AND NOTES


(1) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for Applying the Quantitative Structure−Activity Relationship Paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.

(2) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.

(3) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.

(4) Maggiora, G. M. On Outliers and Activity Cliffs − Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(5) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a Virtual High Throughput Screen by 4D-QSAR Analysis: Application to a Combinatorial Library of Glucose Inhibitors of Glycogen Phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160.

(6) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(7) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview over the Method and Applications, including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(8) Bajorath, J. Virtual Screening: Methods, Expectations, and Reality. *Curr. Drug Discovery* **2002**, *2* (3), 24−28.

(9) Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. Molecular Similarity Analysis and Virtual Screening in Binary-Transformed Chemical Descriptor Spaces with Variable Dimensionality. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 21−29.

(10) Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC − A Virtual Screening Method for the Identification of Potent Hits. *J. Med. Chem.* **2004**, *47*, 4286−4290.

(11) Godden, J. W.; Bajorath, J. A Distance Function for Retrieval of Active Molecules from Complex Chemical Space Representations. *J. Chem. Inf. Model.* **2006**, *46*, 1094−1097.

(12) Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median Partitioning: A Novel Method for the Selection of Representative Subsets from Large Compound Pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885−893.

(13) The Penn Center for Molecular Discovery (PCMD). http://www.seas.upenn.edu/∼pcmd/ (accessed Jan 2007).

(14) AID 453 − PubChem BioAssay Summary. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=453#aLinks (accessed Jan 2007).

(15) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(16) *MACCS Structural Keys*; MDL Elsevier: San Leandro, CA. http://www.mdl.com (accessed Sep 1, 2006).

(17) *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc.: Montreal, Canada. http://www.chemcomp.com (accessed Feb 1, 2006).

(18) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(19) Kubinyi, H. Similarity and Dissimilarity − A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *11*, 225−252.

(20) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(21) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(22) Parker, C. N. McMaster University Data-Mining and Docking Competition: Computational Models on the Catwalk. *J. Biomol. Screening* **2005**, *10*, 647−648.

(23) Parker, C. N.; Shamu, C. E.; Kraybill, B.; Austin, C. P.; Bajorath, J. Measure, Mine, Model, Manipulate: The Future for HTS and Chemoinformatics? *Drug Discovery Today* **2006**, *11*, 863−865.

(24) Pavlidis, P.; Nobel, W. S. Matrix2png: A Utility for Visualizing Matrix Data. *Bioinformatics* **2003**, *19*, 295−296.


CI6005432