Similarity Search Profiles as a Diagnostic Tool for the Analysis of Virtual Screening Calculations

Ling Xue,[†] Jeffrey W. Godden,[†] Florence L. Stahura,[†] and Jürgen Bajorath*[,†,‡]

Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011-8012, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received January 30, 2004

An analysis method termed similarity search profiling has been developed to evaluate fingerprint-based virtual screening calculations. The analysis is based on systematic similarity search calculations using multiple template compounds over the entire value range of a similarity coefficient. In graphical representations, numbers of correctly identified hits and other detected database compounds are separately monitored. The resulting profiles make it possible to determine whether a virtual screening trial can in principle succeed for a given compound class, search tool, similarity metric, and selection criterion. As a test case, we have analyzed virtual screening calculations using a recently designed fingerprint on 23 different biological activity classes in a compound source database containing ∼1.3 million molecules. Based on our predefined selection criteria, we found that virtual screening analysis was successful for 19 of 23 compound classes. Profile analysis also makes it possible to determine compound class-specific similarity threshold values for similarity searching.

INTRODUCTION

Similarity searching using bit string representations of molecular structure and properties, so-called fingerprints, is a computational technique that is widely used in database analysis and pharmaceutical research.[1] Such fingerprints can substantially vary in their length, design, and type of molecular descriptors they capture.[2,3] Fingerprint-based similarity searching is also among the preferred techniques for virtual screening (VS) of large compound databases, for which rather diverse computational methodologies have been adapted.[3] VS efforts typically aim at the identification of novel active molecules, which challenges similarity-based methods to go beyond the detection of structural similarity and extrapolate from the analysis of molecular structures and properties to the recognition of similar biological activities.

Fingerprint-based VS requires the calculation of fingerprints for both query and database compounds and a quantitative comparison of fingerprint overlap as a measure of molecular similarity.[1] A variety of similarity metrics are available for bit string comparisons,[1,4] with the Tanimoto coefficient (Tc) being most widely used. For quantitative comparison, it is also required to establish threshold values for any chosen coefficient or metric as an indicator of true similarity. This is often a rather challenging task since appropriate threshold values depend on the chosen fingerprints and also on the studied compound classes.[2,5] Accordingly, only a few studies have thus far systematically addressed these issues and established (at least approximate) similarity threshold values for specific fingerprint designs,[6−9] although it is in general not possible to firmly associate these values with specific biological activities.[5] Therefore, scaling and data fusion techniques as well as consensus scoring schemes have been developed to further increase the predictive performance of similarity searching.[9−11]

Despite the efficiency of many algorithms and computational screening tools, VS analysis becomes a significant task in many instances, due to the dramatic growth of compound source databases. Simply put, searching for active compounds such as needles in haystacks is in general a nontrivial problem but further complicated by the rapidly growing size of haystacks, which increases the probability of detecting false-positives and missing potential hits. Although a number of studies have been reported that retroactively determined the success of virtual screens on a case-by-case basis or enhancements in hit rates once experimental data became available,[3,12] it is rather difficult to draw from such case studies general conclusions about the performance of VS tools or the probability to succeed with a specific application.[5,13]

We are interested in approaches that enable us to look at the potential of VS calculations from a more principal point of view, for example, by trying to understand whether a VS trial could in principle be successful, given certain compound classes, selection criteria, and search tools. Here we describe a study focusing on fingerprint-based VS that was designed to address such questions. We systematically tested a recently introduced novel type of fingerprint, MP-MFP, on a total of 23 classes of compounds with different activities. Data analysis was facilitated by generation of similarity search

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com. Corresponding author address: Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, WA 98011-8012.

[†] AMRI Bothell Research Center (AMRI-BRC).

[‡] University of Washington.

10.1021/ci040120g CCC: $27.50
Published on Web 05/05/2004

Table 1. Activity Classes and Search Results[a]

abbr	biological activity	*N*	$[avTc]_I$	$[hit/DBC]_I$	hits/100DBC	$[avTc]_{100}$
BLC	β-lactamase inhibitors	7	0.66	4.9/162.7	4.9	0.67
PKC	protein kinase C inhibitors	7	0.71	2.9/64.6	2.9	0.70
ADR	adrenergic receptor ligands	8	0.69	6.3/191.6	5.5	0.72
GLU	glucocorticoid analogues	7	0.83	1.4/45.3	2.0	0.80
BEN	benzodiazepine receptor ligands	11	0.76	2.9/43.2	3.3	0.75
CAE	carbonic anhydrase II inhibitors	11	0.73	2.9/50.9	3.1	0.72
H3E	H3 antagonists	10	0.69	7.2/192.7	7.0	0.71
TKE	tyrosine kinase inhibitors	10	0.82	1.2/25.5	1.8	0.79
5HT	serotonin receptor ligands	10	0.81	1.2/28.4	1.6	0.77
HIV	HIV protease inhibitors	9	0.75	1.3/39.1	1.8	0.74
COX	cyclooxygenase-2 inhibitors	8	0.76	2.3/58.9	2.3	0.75
ANG	angiotensin AT1 antagonists	5	0.70	2.0/81.2	2.0	0.70
ARO	aromatase Inhibitors	5	0.78	0.4/14.8	0.4	0.75
DIH	dihydrofolate reductase inhibitors	5	0.77	0.8/36.2	0.8	0.75
FAC	factor Xa inhibitors	7	0.76	1.1/38.4	1.1	0.74
MAT	matrix metalloproteinase inhibitors	6	0.69	3.3/141.7	3.3	0.69
VIT	vitamin D analogues	6	0.72	4.3/153.0	4.3	0.73
RTI	reverse transcriptase inhibitors	7	0.76	0.6/16.9	0.9	0.72
PPAR	PPARgamma agonists	8	0.69	3.8/88.4	3.8	0.70
DD2A	dopamine D2 antagonists	7	0.72	0.9/28.9	1.4	0.70
CRF1	CRF1 antagonists	6	0.75	1.0/41.5	1.3	0.73
CALA	calcium antagonists	9	0.86	0.2/5.0	0.9	0.77
ARI	aldose reductase inhibitors	6	0.80	0.3/11.0	1.0	0.73

[a] For each activity class, the total number (*N*) of compounds is reported (accordingly, in our calculations, the number of potential hits is ($N-1$)). $[avTc]_I$ is the similarity threshold value at the intersection point in SSPs and $[hit/DBC]_I$ the ratio of correctly identified hits and detected background compounds at this point; hits/100DBC reports the number of hits contained in ~100 selected (active plus database) compounds and $[avTc]_{100}$ the corresponding similarity threshold value.

profiles (SSPs) that we designed to monitor compound recognitions during VS as a function of similarity threshold values. This intuitive graphical approach has made it possible to determine the suitability of fingerprint searches on various compound classes and study general trends with relevance for compound selection. In addition, the findings reported herein also support our view that MP-MFP represents an attractive similarity search tool.

METHODS

Fingerprint. MP-MFP[14] belongs to a series of relatively short fingerprints (so-called minifingerprints or MFPs) that combine specifically selected molecular property and structural fragment-type descriptors.[8,15] It consists of a total of 175 bit positions, 61 of which correspond to binary encoded molecular property descriptors.[14] The presence of binary encoded property descriptors distinguishes MP-MFP from other fingerprint designs.[14,15] Binary transformation of descriptors with continuous value ranges is facilitated by calculation of statistical medians of their value distributions in large compound databases.[16] Thus, a bit assigned to a binary transformed descriptor is set to one if its value for a compound is equal to or greater than the median and to zero if it is smaller. In a number of initial test calculations, MP-MFP was compared to the publicly available set of 166 MACCS keys[17] and other MFPs and found to perform better overall.[14] An SVL-based implementation[18] of MP-MFP is publicly available.[19]

Similarity Metric. As a similarity coefficient, a previously introduced Tc variant was applied termed average Tc (avTc).[14]

Tc is defined as $Tc = bc/(b1+b2-bc)$, with b1 being the number of bits set in molecule 1, b2 the number of bits set in molecule 2, and bc the number of bits set in common with both molecules.

avTc is defined as $avTc = (Tc+Tc')/2$, with Tc′ being the Tc calculated for bit positions set to zero, rather than to one (as in conventional Tc calculations).

For similarity searching with MP-MFP, the introduction of avTc was required because for binary transformed descriptors, as described above, bit positions set to zero capture as much information as those set to one and must therefore be taken into account when quantifying fingerprint overlap.

Test Compounds. As search templates and potential hits, 175 compounds belonging to 23 different activity classes were used that were originally assembled for fingerprint scaling analysis.[9] The composition of this data set is summarized in Table 1. Active compounds were added to a database containing ~1.3 million molecules collected from various vendor sources[16] that served here as our source for VS calculations.

Calculations. For each activity class, one compound at a time was searched against the source database after adding the remaining active compounds as potential hits. For example, for activity class BLC in Table 1, each of the seven compounds was used once as a search template and in each case six potential hits with similar activity were present in our large source database. During each trial, avTc threshold values were systematically varied from zero to one in 0.01 increments. Any database compound other than potential hits that was detected as similar at a given threshold value was thought to be inactive for the purpose of this analysis (although some of these background molecules might well be hits). Routines required for all VS calculations were implemented in the Molecular Operating Environment.[20]

Similarity Search Profiles. For profile generation, correctly identified hits and other detected background molecules were separately monitored at each avTc threshold value, and the results were averaged over all compounds belonging to an activity class. Active and background

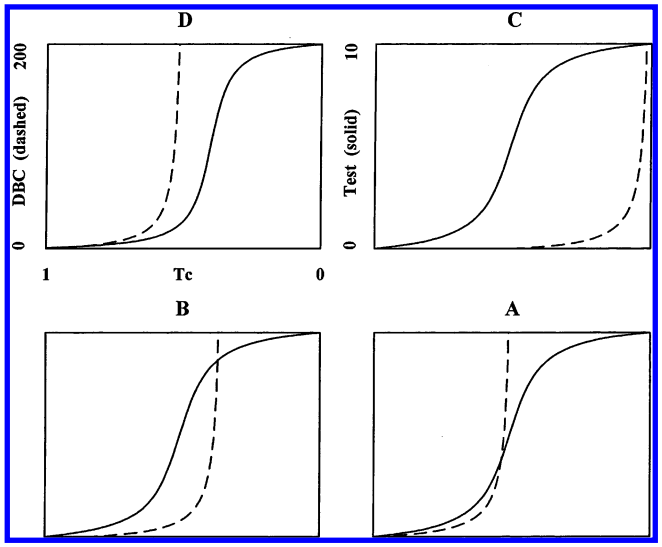


Figure 1. Schematic similarity search profiles. Model SSPs are shown that illustrate the basic idea of profile analysis and are indicative of unsuccessful or successful VS trials, as discussed in the text. Test stands for test compounds (i.e., 0−10 potential hits) and DBC for (0−200) database compounds.

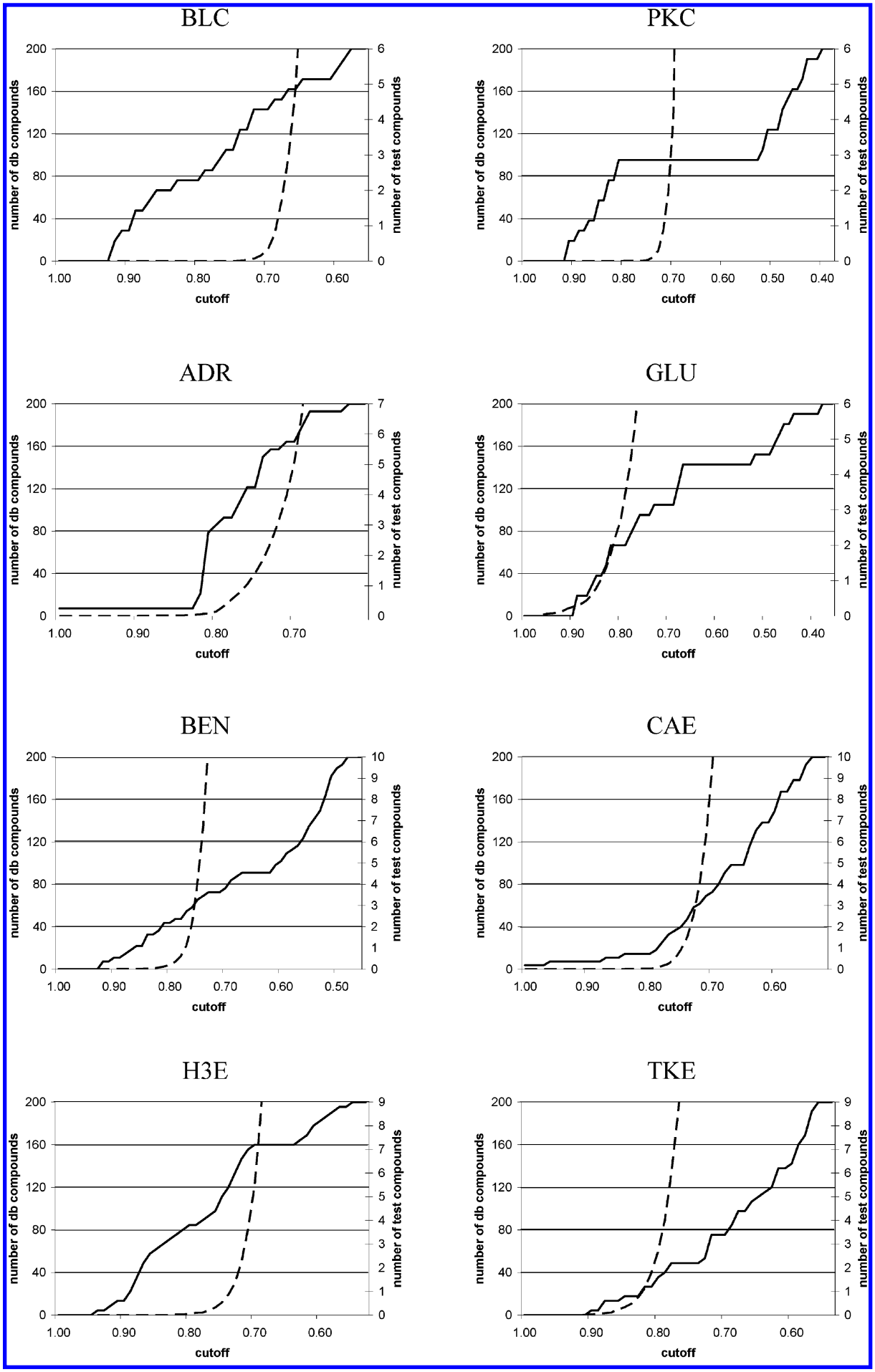
molecules were presented on the same plot over the entire avTc range using separate scales. Model SSPs are shown in Figure 1. Solid and dashed lines represent the average of correctly identified hits and background molecules, respectively. To obtain a diagnostic tool to evaluate the potential and limitations of VS calculations, the range of database compounds monitored is determined based on appropriate compound selection criteria, as further rationalized in the Results section. Thus, SSPs capture a window of the overall VS trial that focuses on ranges of hits and background compounds and their ratios that correspond to preferred results. This type of profiling technique is distinct from other graphical methods such as cumulative recall curves[21] that have been used, for example, to monitor hit rates in similarity search calculations.[13,21]

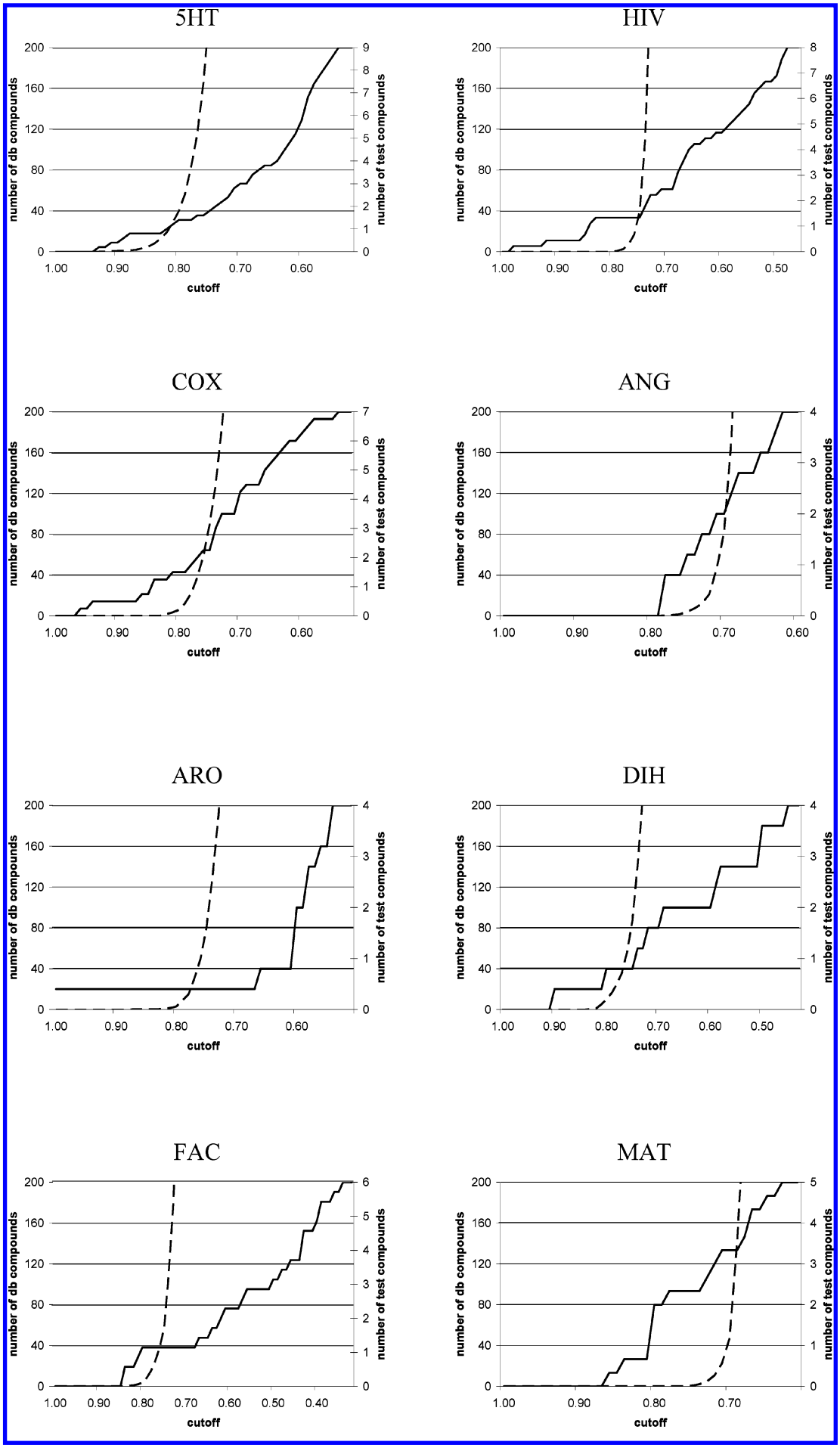
RESULTS AND DISCUSSION

Compound Selection Criteria. In our source database, only a small number of potential hits was present for each activity class (between five and ten) but a large number of background compounds (∼1.3 million), which represents a challenging scenario for VS calculations. In our analysis, we intended to mimic a practical VS situation as closely as possible. Thus, a reasonably small number of database compounds had to be selected (as putative candidates for testing). Therefore, in each case, we focused on selecting ∼100 compounds, a fairly typical number of candidates. Our minimum criterion for a successful VS trial was that the selected molecules needed to contain on average at least one active compound per activity class (corresponding to a 1% hit rate).

Profile Generation and Analysis. The model profiles in Figure 1 cover the range of preferred VS results according to our selection criteria. In this example, zero to 200 background compounds are monitored and zero to 10 potential hits. As the similarity threshold value and the corresponding search stringency increase during VS calculations, fewer and fewer compounds are detected as similar to the template molecules. Due to the different scales on the compound (ordinate) axes, all of the hits are monitored, but the curve representing the database compounds enters the profile window only at a certain level of search stringency (i.e., if fewer than 200 background molecules are recognized). The best possible indicator of a successful outcome of the VS trial is the presence of an intersection point of the hit and database compound curves within the profile window. For example, in Figure 1A, the intersection point indicates that ∼100 database compounds contain three to four hits, a desired result. Similarly, in Figure 1B, at the intersection point 180−190 database molecules would contain eight or nine hits, and, extrapolated from there, 100 database compounds contain essentially the same number of hits. Thus, a trial yielding this profile would also be successful according to our criteria. In Figure 1C, the profile corresponds to an almost ideal outcome of a VS experiment because the chosen fingerprint detects the majority of hits at stringency levels where only very few, if any, other database compounds are recognized. However, findings of this nature should be cautiously evaluated as they could well indicate that the chosen fingerprint specifically responds to features only shared by a class of active compounds. Thus, this search tool might be "overtrained" and not more generally applicable, which could be confirmed by studying other compound classes. By contrast, in Figure 1D, the graph represents VS calculations that cannot possibly succeed and reveals reasons for this. Here the particular fingerprint detects by far too many background molecules as similar to the templates and only sufficiently reduces the number of recognized compounds at very high levels of search stringency where even true similarity between templates and hits can no longer be detected. Such findings would indicate that the fingerprint might be much too insensitive to molecular features determining biological activity. Thus, analysis of SSPs makes it possible to better understand why VS calculations succeed or fail.

Virtual Screening Trials. In Figure 2, SSPs are shown for each of the 23 activity classes tested here in systematic VS calculations. As can be seen, most of the profiles display intersection points within the desired range, thus indicating successful trials. Depending on the particular compound class, SSPs show in part significant differences, illustrating that VS results are generally influenced by compound class-specific features. However, it is easy to see that activity classes such as BLC, H3E, or MAT represent rather successful cases. A quantitative analysis of the profiles is reported in Table 1 and Figure 3 illustrates the analysis of a representative profile. Based on the data in Table 1, the influence of compound class-specific features on SSPs is further illustrated by differences in avTc values at intersection points that range from 0.66 to 0.86. The ratio of correctly identified hits versus other database compounds at intersection points provides a quantitative measure of VS performance. For example, for activity class PKC the detection of on average three active and 65 database compounds corresponds to an intersection point hit rate of 4.5%. Furthermore, from the intersection point, it is straightforward to extrapolate to the level of ∼100 database compounds (as illustrated in Figure 3) and calculate hit rates in accordance with our selection criteria. Only for four of 23 classes (ARO, DIH, RTI, and CALA) an average of less than one hit per 100 database molecules was observed, a VS failure according to our criteria. Inspection of the SSPs of these classes revealed


BLC
number of db compounds
number of test compounds
cutoff
PKC
number of db compounds
number of test compounds
cutoff
ADR
number of db compounds
number of test compounds
cutoff
GLU
number of db compounds
number of test compounds
cutoff
BEN
number of db compounds
number of test compounds
cutoff
CAE
number of db compounds
number of test compounds
cutoff
H3E
number of db compounds
number of test compounds
cutoff
TKE
number of db compounds
number of test compounds
cutoff


5HT
number of db compounds
number of test compounds
cutoff
HIV
number of db compounds
number of test compounds
cutoff
COX
number of db compounds
number of test compounds
cutoff
ANG
number of db compounds
number of test compounds
cutoff
ARO
number of db compounds
number of test compounds
cutoff
DIH
number of db compounds
number of test compounds
cutoff
FAC
number of db compounds
number of test compounds
cutoff
MAT
number of db compounds
number of test compounds
cutoff

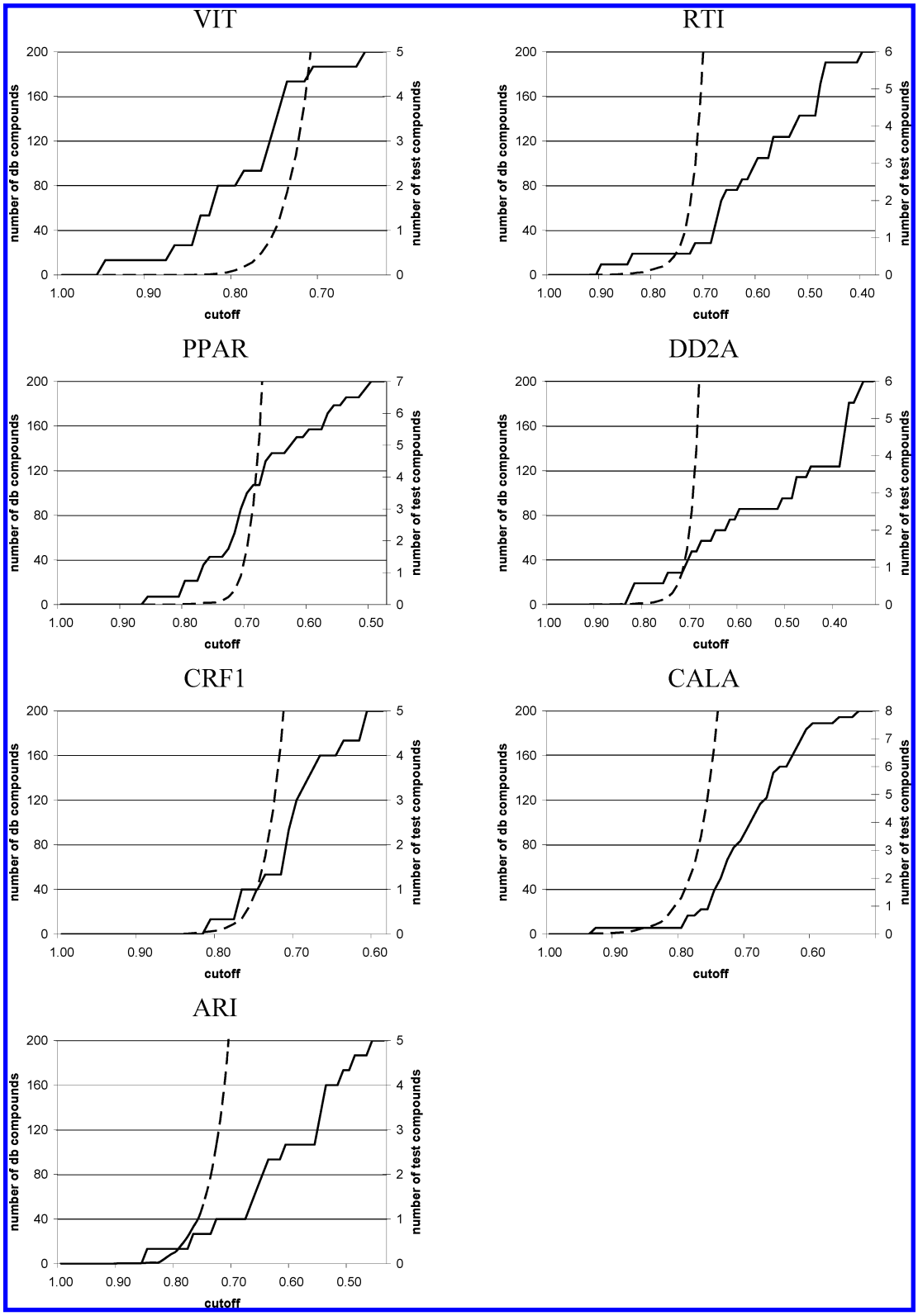


Figure 2. Similarity search profiles for diverse compound classes. Shown are SSPs capturing VS trials on 23 different activity classes. In contrast to the curve for the database compounds (or idealized curves shown in Figure 1), the curve representing test compounds (potential hits) is not smooth because only a few active molecules are monitored. Activity classes are abbreviated as in Table 1. On the horizontal axis, cutoff designates avTc similarity threshold values.

that in all four cases this failure could essentially be attributed to the situation discussed for the model profile in Figure 1D, i.e., too little sensitivity of the fingerprint toward a compound class. However, such observations were exceptions in this study. For the successfully searched compound classes, best observed hit rates ranged from 5% to 7%.

Given the well-documented fact that different similarity search tools often perform differently dependent on the specifics of compound classes and search situations,[2,13] the overall performance of MP-MFP was considered encouraging, since the test calculations gave at least satisfactory results for 19 of 23 activity classes studied here. However, the most

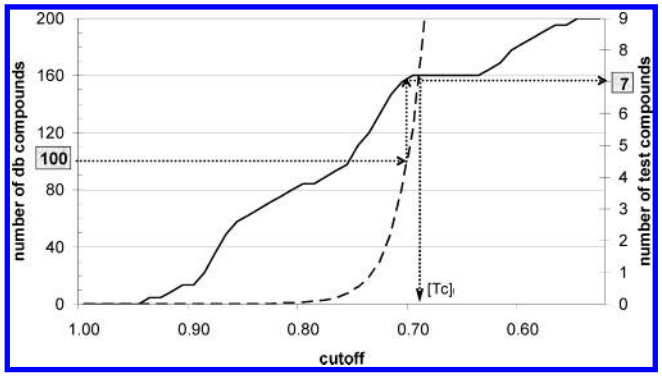


Figure 3. Analysis of similarity search profiles. As an example, the profile of activity class H3E is shown. The figure illustrates how similarity threshold value and compound numbers reported in Table 1 are determined.

important aspect is that SSP analysis was designed to study and compare different similarity search methods and answer performance questions from a more principal point of view. Thus, the approach does not depend on the application of specific search tools.

Similarity Threshold Values. We also found that extrapolation from intersection points to the level of ~100 database compounds narrowed the range of avTc threshold values, as shown in Table 1. With only two exceptions, all $[avTc]_{100}$ values fell within the range between 0.70 and 0.80. These results indicate that the avTc interval [0.70, 0.80] is a preferred similarity threshold range for MP-MFP, which has implications for practical VS applications. Analysis of single search calculations indeed revealed that encouraging results could be obtained by making use of these findings. When we searched each active compound in successfully treated classes at a fixed avTc = 0.80 against the source database, as one would need to do without any prior knowledge of compound class-specific threshold values, between one and six hits were identified in 107 of 142 calculations (75%). However, the number of database compounds detected in these calculations was typically much smaller than 100 (often 10 or even fewer), and only in three cases were more than 100 compounds detected. Thus, in calculations under relatively stringent similarity threshold conditions, false-positive numbers could often be further reduced, while retaining significant hit rates, at least in this case.

Given the general applicability of SSP analysis, our findings suggest that it might often be possible to identify preferred ranges of chosen similarity coefficients and fingerprints based on a detailed analysis of search profiles. This could be helpful in tuning similarity search calculations toward specific compound classes and desired selection criteria.

CONCLUSIONS

In this study, we have introduced a graphical approach to analyze fingerprint-based VS calculations in detail and determine whether these trials could succeed. As described, the method can also be applied to evaluate the performance of similarity search tools other than fingerprints, as long as quantitative similarity measures are used. Moreover, the analysis of SSPs can serve not only as a diagnostic but also as a predictive tool, which has practical implications for VS investigations. For example, SSPs can be calculated whenever a number of different template compounds with similar activity are available as a starting point for VS, which is often the case. Based on the resulting profiles, it can then be estimated what the probability of a successful search for novel active compounds might be, given the specific composition of the source database and the chosen similarity search tool(s). For example, if searches among known actives produce unfavorable SSPs in a given source database, the probability of identifying novel hits is likely to be low. Based on the results obtained for the test calculations reported herein, we can also conclude MP-MFP produces promising VS results over a broad range of compound classes. In addition, it has been possible to determine a preferred similarity threshold range for its application.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(2) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(3) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **2002**, *1*, 882−894.
(4) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155−166.
(5) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(6) Patterson, D. E., Cramer, R. D., III; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(7) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(8) Xue, L.; Godden, J. W.; Bajorath J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.
(9) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218−1225.
(10) Salim, N.; Holliday, J. D.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.
(11) Wang, R.; Wang; S. How does consensus scoring work for virtual library screening? An idealized computer experiments. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(12) Engels, M. F. M.; Venkatarangan, P. Smart screening: approaches to efficient HTS. *Curr. Opin. Drug. Discov. Dev.* **2001**, *4*, 275−283.
(13) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903−911.
(14) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.
(15) Xue, L.; Godden, J. W.; Bajorath, J. Mini-fingerprints for virtual screening: design principles and generation of novel prototypes based on information theory. *SAR QSAR Environ. Res.* **2003**, *14*, 27−40.
(16) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive median partitioning for virtual screening of large databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182−188.
(17) MACCS keys. MDL Information Systems Inc., San Leandro, CA. URL: www.mdl.com.
(18) Sanatvy, M.; Labute, P. SVL: the scientific vector language. *J. Chem. Comput. Group*. URL: www.chemcomp.com/feature/svl.htm.
(19) SVL Exchange. URL: svl.chemcomp.com.
(20) Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, Quebec, Canada. URL: www.chemcomp.com.
(21) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *J. Mol. Graph. Model.* **2000**, *18*, 343−357.

CI040120G