Sharing Chemical Information without Sharing Chemical Structure

Brian B. Masek,*,‡ Lingling Shen,†,§ Karl M. Smith,‡ and Robert S. Pearlman†

Laboratory for the Development of Computer-Assisted Drug Discovery Software, College of Pharmacy, University of Texas at Austin, Austin, Texas 78712, and Tripos, Inc., 1699 S. Hanley Road, St. Louis, Missouri 63144

Received August 31, 2006

Studies to assess the risks of revealing chemical structures by sharing various chemical descriptor data are presented. Descriptors examined include "Lipinski-like" properties, 2D-BCUT descriptors, and a high-dimensional "fingerprint-like" descriptor (MACCs-vector). We demonstrate that unless sufficient precautions are taken, de novo design software such as EA-Inventor is able to derive a unique chemical structure or a set of closely related analogs from some commonly used descriptors. Based on the results of our studies, a set of guidelines or recommendations for safely sharing chemical information without revealing chemical structure is presented. A procedure for assessing the risk of revealing chemical structure when exchanging chemical descriptor information was also developed. The procedure is generic and can be applied to any chemical descriptor or combination of descriptors and to any set of structures to enable a decision about whether the exchange of information can be done without revealing the chemical structures.

INTRODUCTION

"Safely" exchanging chemical information has recently received increased attention in the QSAR and CADD communities.[1,2] There are several reasons for this interest. The primary motivation concerns the need to develop general predictive models for ADME, toxicology, and other biological and physical properties, which continues to be a challenging area of research. There is considerable interest in improving both the chemical scope and predictive power of models based on measured properties from ever larger and more diverse sets of compounds. Pooling and sharing measured property and/or activity data between research groups could greatly increase the predictive power of QSARs and QSPRs for the mutual benefit of the cooperating groups and for the benefit of the scientific community as a whole. However, since much of this measured information is generated by proprietary industrial pharmaceutical and agrochemical research efforts, a large pool of relevant information has not been disclosed in the public domain. Private sector companies remain very reluctant to reveal information which could enable a competitor to deduce the chemical structure of proprietary compounds for fear of damaging intellectual property rights or losing competitive advantage.

A second motivation for exchanging information comes from efforts to enhance pharmaceutical or agrochemical compound collections or screening libraries. For example, a pharmaceutical company might be interested in acquiring or testing compounds with certain properties and a technology partner might be interested in providing compounds but does not wish to disclose the structures of their proprietary compounds prior to the sale. Again, the issue is how to "safely" exchange relevant chemical information without revealing chemical structure.

In this paper, we describe how appropriate software can be used to assess the risk that chemical structure may inadvertantly be revealed by sharing chemical descriptors.[3] We have used the EA-Inventor software to generate structures de novo that have the same chemical descriptor values as a target compound and to study the ability to deduce a specific chemical structure from various types of chemical descriptors. The procedures presented are generic and could be applied to any chemical descriptor or combination of descriptors and to any set of structures to enable a decision about whether the exchange of information can be done without indirectly revealing the chemical structures.

Our studies show that unless sufficient precautions are taken, EA-Inventor is able to discover target structures or closely related analogs from some types of commonly used descriptors. Based on these results, a set of guidelines or recommendations for safely sharing chemical information are presented.

METHODS

For deducing or "inventing", suitable de novo design software is required to deduce or "invent" chemical structures that have the same descriptor values as a given target compound.[4] Such software needs to be able to work with whatever types of descriptor the user wishes to study and to drive the evolution of invented structures using those descriptors. EA-Inventor was designed to work with external scoring functions that make use of awk, perl, or shell scripts which evaluate a single scoring function or some combination of scoring functions and filters. At each generation, the structures to be scored are written out to a file. The program then runs the user-specified script to generate scores. For the studies reported in this paper, simple c-shell scripts were

* Corresponding author phone: (314)951-3409; fax: (314)647-9241; e-mail: bmasek@tripos.com.
† University of Texas at Austin.
‡ Tripos, Inc.
§ Current address: Pfizer Inc., 620 Memorial Drive, Cambridge, MA 02139.

10.1021/ci600383v CCC: $40.75
Published on Web 02/07/2008

written to perform the specified descriptor calculations, to generate the scores, and to write the scores to a file. Scores are then read by EA-Inventor and used to guide the next iteration of the evolutionary process.

Scores were calculated as follows. Let *t* be a target compound for which descriptors are provided and a structure is to be deduced. Let *i* be any of the structures generated during the process of trying to deduce the structure of the target. The scoring function, *D*, is the Euclidian distance given by the equation

$$D_{t,i} = \sqrt{\sum_{j=i}^{N_d} (d_t - d_i)^2}$$

where N_d is the number of descriptors.

EA-Inventor v4.1[5] was used for all of the studies reported here. Default settings were used unless otherwise noted. A population size of 100 structures was used for all EA-Inventor runs. ClogP was calculated with the BioByte ClogP[6] program. BCUT descriptors and MACCs counts vectors were calculated using DiverseSolutions v6.2.[7−9] All other descriptors were calculated using StructureFilter v1.0.1.[5] All studies were performed on either an SGI Origin 2000 system or SGI Indigo workstations. The CPU time required for a study of a single target varied from a few minutes to 1−2 h, depending on the descriptors used. The CPU time consumed was dominated by the time required to calculate descriptor values; the CPU time consumed by EA-Inventor was minor in most cases.

Stereochemistry. There is no way for EA-Inventor to decipher the stereochemistry of the target from the 2D descriptors considered here. Stereochemistry was therefore not a consideration when assessing the distance, D_{ti}, between invented structures and the target compounds. EA-Inventor's "ScoreIs2D" option was used, with the result that atom and bond stereochemistry was left unspecified in the structures generated.

Overview of the Assessment Procedure. The following assessment procedure was used to determine whether it is possible, using appropriate software, to deduce chemical structure from chemical descriptors.

1. A diverse set of 100 "target structures" to be used in all tests was selected from the MDL Drug Data Report (MDDR) database[10] using the distance-based diverse subset selection algorithm in DiverseSolutions.

2. In order to explore the notion that reduced descriptor precision prevents the deduction of chemical structure from descriptors, various levels of precision were selected for descriptors that were not integer-valued.

3. For each descriptor (or combination of descriptors) and for each precision level, we attempted to deduce each target structure using EA-Inventor to invent structures which have the same descriptor value(s) as the target structure.

   a. For each of the 100 target structures, an initial population of 100 structures to start the evolutionary process used by EA-Inventor was chosen from MDDR based on the "descriptor distance" (see 3b) to the target compound. If a structure was found to be identical to the target (ignoring atom and bond stereochemistry), it was replaced. Our premise was that anyone attempting to deduce a structure from descriptors would likely start by searching a database for structures with descriptor values as close as possible to the descriptors of the target structure.

   b. For each set of descriptors and for each invented structure *i* we computed the Euclidean distance D_{ti} between the invented structure and the target structure *t* based on the descriptors being considered. This distance was used as the "score" in the EA-Inventor evolutionary process for inventing structures.

   c. We terminated EA-Inventor processing if all 100 structures in the invented population had a zero score (zero distance) or after 500 generations, whichever came first. There were two reasons we allowed invention to continue after the first structure with zero distance to the target molecule had been found. First, if the first invented "0-score" structure found was the same as the target structure, we would not know if EA-Inventor might have subsequently found some other structure with a zero score. Second, we wished to determine how uniquely a structure can be determined by generating a set of structures with the same descriptor value(s).

4. For each target structure (and for each set of descriptors), we noted whether any 0-score structures were invented, and, if so, whether any of them were actually the same as the target structure.

Table 1. Descriptors

descriptor	summary
MW	molecular weight
N_{Hb-D}	number of hydrogen bond donors
N_{Hb-A}	number of hydrogen bond acceptors
N_{rot}	number of rotatable bonds
N_{SCA}	number of atoms with chirality explicitly specified
MlogP	Moriguchi LogP
ClogP	ClogP
(nth)2D-BCUT	one of the five coordinates in the five-dimensional 2D-BCUT chemistry space
(1−5)2D-BCUT	all five coordinates in the five-dimensional 2D-BCUT chemistry space
MACCs-vector	a vector whose 166 elements are the number of occurrences of substructural features used as MACCs database keys[12]

RESULTS

The nine commonly used descriptors chosen for study are listed in Table 1. A number of "Lipinski-like" properties were studied, since they are often used in predictive modeling. We also chose to study BCUT descriptors which are widely used for chemical diversity, library design, and QSAR studies. Five 2D-BCUT descriptors[11] were selected using DiverseSolutions. Those chosen best represent the diversity of the MDDR database, a database composed of druglike compounds. To determine to what degree each descriptor reveals chemical structure, the BCUT and Lipinski-like descriptors were studied individually. The results for these studies are given in Table 2. Various combinations of these descriptors were also examined in hopes of gaining insight into whether some such combinations reveal more details of chemical structure than other combinations or individual descriptors do. We also chose to study a high-dimensional "fingerprint-like" descriptor (MACCs-vector). The 166 substructural fragments used by MDL's MACCs program as

Table 2. Summary of Studies To Deduce the Structures of 100 Targets from Single Descriptors

study no.	descriptor[a]	targets with 0-Score[b]	structures with 0-Score[c]	invented equals target[d]
1	MlogP (3)	100	10000	6
2	ClogP (3)	100	10000	7
3	MW (6)	62	6134	6
4	N_{Hb-D} (1)	100	10000	3
5	N_{Hb-A} (1)	100	10000	1
6	Nrot (1)	100	10000	3
7	N_{SCA} (1)	100	10000	2
8	first 2D-BCUT (3)	100	10000	8
9	second 2D-BCUT (3)	100	10000	9
10	third 2D-BCUT (3)	100	10000	6
11	fourth 2D-BCUT (3)	100	10000	11
12	fifth 2D-BCUT (3)	100	10000	4
13	first 2D-BCUT (5)	100	9736	13
14	second 2D-BCUT (5)	100	10000	13
15	third 2D-BCUT (5)	100	10000	11
16	fourth 2D-BCUT (5)	100	9964	13
17	fifth 2D-BCUT (5)	100	10000	9
18	first 2D-BCUT (7)	96	1205	16
19	second 2D-BCUT (7)	88	5832	17
20	third 2D-BCUT (7)	96	4838	15
21	fourth 2D-BCUT (7)	98	2352	14
22	fifth 2D-BCUT (7)	100	10000	16

[a] The number in parentheses after the descriptor name is the "precision" (the number of significant figures used). [b] The number of targets for which EA-Inventor was able to invent a structure with the same descriptor value as the target molecule. [c] The number of structures that EA-Inventor was able to find with the same descriptor value as the target molecule. The maximum number of "0-scores" possible is 10 000 = 100 (population size) × 100 targets. [d] The number of target structures found (maximum = 100).

database keys[12,13] were used to generate a vector whose elements are the number of occurrences or "counts" of each substructure. These substructural fragments were intended for use in substructure and exact structure searching and were chosen to optimize structure retrieval. While collision rates in database searching for the binary fingerprint based on these fragments have been well studied,[14−16] the use of structural counts vector is less well studied. Results from studies with combinations of descriptors and high-dimensional descriptors are summarized in Table 3.

DISCUSSION

Individual Descriptors. We expected that none of these descriptors, individually, would contain sufficient information to uniquely deduce chemical structure. The results of our studies confirmed this expectation (Table 2). Several details merit discussion, however. The number of "Structures with 0-Score" found by EA-Inventor gives an indication of whether a descriptor contains sufficient information to uniquely determine the chemical structure. In most of the single descriptor studies, EA-Inventor finds the maximum number of 0-score structures possible: 10 000 = 100 (population size) × 100 targets. In these cases EA-Inventor has found, for each target, 100 *different* structures with the same descriptor value as the target. In all of the other single descriptor studies, EA-Inventor is able to identify *many* structures with the target descriptor value. Occasionally, the actual target structure is found, but, in all cases, it is only one of the *many* 0-score structures found. Whether we generate the actual target structure (with 0-score) or other

Table 3. Summary of Studies To Deduce the Structures of 100 Targets from Combinations of Descriptors

study no.	descriptor[a]	targets with 0-Score[b]	structures with 0-Score[c]	invented equals target[d]
23	(1−7)[e] 1D	34	2198	19
24	(1−5) 2D-BCUT (7)	33	33	33
25	(1−5) 2D-BCUT (5)	32	32	32
26	(1−5) 2D-BCUT (3)	66	4978	29
27	(1−5) 2D-BCUT (3) plus (1−7) 1D	34	88	34
28	MACCS-vector (0)	79	608	76

[a] The number in parentheses after the descriptor name is the "precision" (the number of significant figures used). [b] The number of targets for which EA-Inventor was able to invent a structure with the same descriptor value as the target molecule. [c] The number of structures that EA-Inventor was able to find with the same descriptor value as the target molecule. The maximum number of "0-scores" possible is 10 000 = 100 targets × 100 (population size). [d] The number of target structures found (maximum = 100). [e] The score was the distance from the target structure in a descriptor space of the seven "Lipinski-like" properties (MlogP, ClogP, MW, N_{Hb-D}, N_{Hb-A}, N_{rot}, and N_{SCA}) from Studies 1−7.

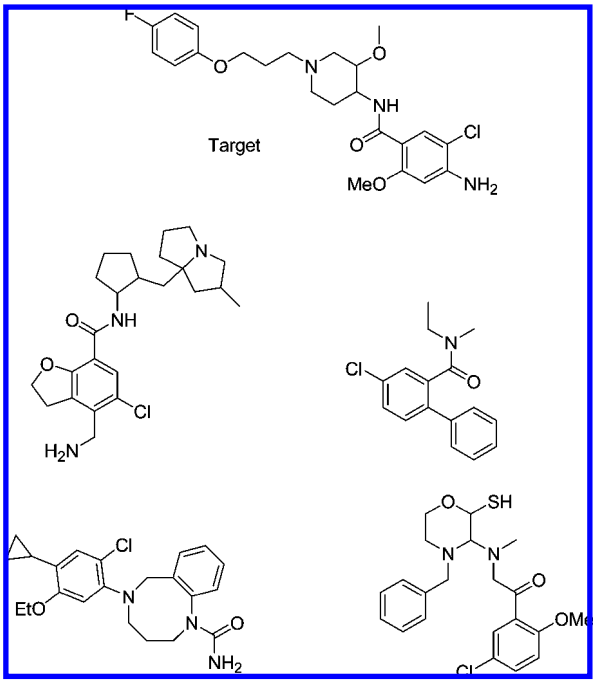

Figure 1. A target structure and four examples of "0-score" structures identified by EA-Inventor in Study 2. All 0-score structures have the same ClogP descriptor value (ClogP = 3.81) as the target structure.

structures (with 0-scores) is merely a matter of luck. It is instructive to look at the 0-score structures generated. Examples from two studies are given in Figures 1 and 2 and indicate that the structures are quite varied. These results show that, as expected, none of the individual descriptors contain sufficient information to allow the chemical structure of the target to be deduced.

If, for every target structure, EA-Inventor is able to invent a structure with the same individual descriptor value (to the precision specified), then the number of "Targets with 0-Score" for that descriptor study will be 100. From Table 2, it can be seen that this is the case for most, but not all, of the studies with individual descriptors. Molecular weight (Study 3) in particular is an example where for a number of

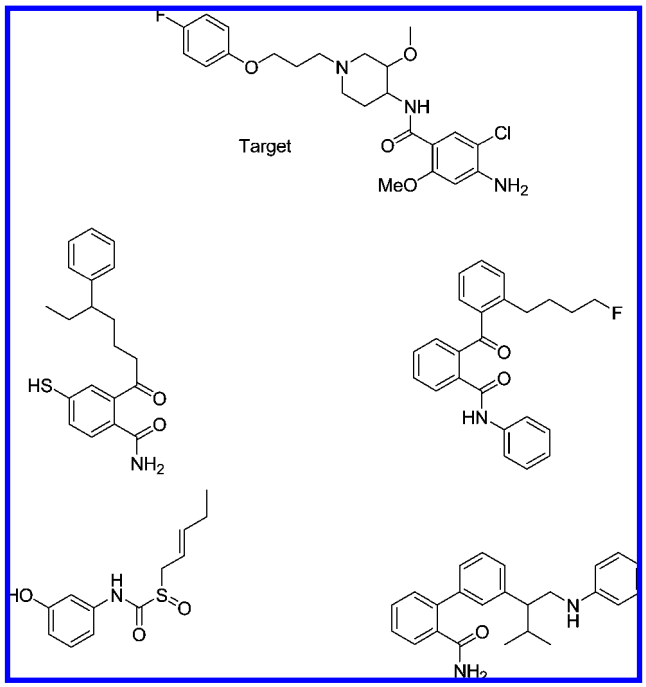

Figure 2. A target structure and four examples of "0-score" structures identified by EA-Inventor in Study 19. All 0-score structures have the same H-bond acceptor 2D-BCUT descriptor values as the target structure. (Note that this descriptor is *not* a simple count of H-bond acceptors but, rather, reflects a combination of the number of such acceptors and their distribution in the structure.)

targets, we fail to find a 0-score structure. Why does EA-Inventor occasionally fail to find a 0-score structure for some targets? The answer is that all evolutionary algorithms are stochastic processes that may sometimes converge to local minima. In these cases, we could improve the results by allowing a larger number of iterations or by using a larger population size or, most effectively, by running the program multiple times with different random seeds.

Combinations of Descriptors and High-Dimensional Descriptors. In contrast to our expectations regarding the inability of individual descriptors to reveal target structures, we thought that some combination of descriptors or a high-dimensional descriptor such as the MACCs-vector descriptor might contain sufficient information to reliably reveal all or at least significant aspects of the chemical structure of the target. Our studies confirmed that expectation (Table 3).

From the summary for Study 23 shown in Table 3, we conclude that, while the combination of Lipinski-like descriptors (MlogP, ClogP, MW, N_{Hb-D}, N_{Hb-A}, N_{rot}, and N_{SCA}) does *not* reveal a unique 0-score structure enabling identification of the target, this combination of descriptors *does* often reveal substructural features of the target. For about a third of the targets, EA-Inventor successfully invents many 0-score structures, indicating that many nontarget structures can be generated with the same Lipinski-like descriptor values. However, by examining the 0-score structures, we observed that the structures contain many of the same substructural elements as their target. Two such examples of sets of 0-Score structures and the corresponding target from Study 23 are shown in Figures 3 and 4. As can be seen, many substructural components of the target can be deduced from examining these structures. The combination of Lipinski-like descriptors contain sufficient information to allow EA-Inventor and a savvy chemist to deduce substructural features of the target structure.

From Study 24, summarized in Table 3, we conclude that the high precision coordinates in a BCUT chemistry space reveal a high level of detail about a chemical structure. Study 24 as well as Studies 25−27 each consists of a single attempt to deduce the structure of a single target. The fact that *only* the target structure was identified as a 0-score structure by EA-Inventor in this single attempt for 33 of the 100 targets implies that a structure *can* be uniquely defined by and deduced from precise coordinates in a BCUT chemistry space. As indicated below, we suspect we could have found the target structure in all (or a much larger fraction) of the other 67 targets if we had allowed EA-Inventor to run for more than 500 generations per target; if we had used a larger population size; or if we had performed multiple attempts using different random seeds and the same target. However, we must note that it is also possible that additional searching might find some nontarget 0-score structures.

From Study 28 we conclude that, while the MACCS-vector descriptor may not uniquely reveal the exact target structure, it does reveal significant details and/or close analogs. Examples of a typical set of 0-Score structures and the corresponding target from this study are shown in Figure 5. As would be expected, the structures found by EA-Inventor are quite similar to the target structure. It is interesting to note that the 166 fragment-counts represented by the MACCS-vector enabled EA-Inventor to find the target structure for 76 of the 100 randomly chosen targets but that these 76 "hits" were "hidden" among 608 − 76 = 532 "false positives." This result stands in contrast to that obtained using the 5 2D-BCUT descriptors (Study 24) in which case EA-Inventor was able to find the target structure for just 34 of the targets, but 100% of these "hits" were exact matches.

In contrast with the "flat" (weak) relationships between the relatively crude, single descriptors and the target structures (Studies 1−22), the strong relationships between structure and the combined or multidimensional descriptors (Studies 23−28) result in very convoluted response surfaces which, ironically, often results in "local minimum problems" and failure to invent the target structure. In order to test this last possibility further, we performed Studies 24a−e wherein we ran EA-Inventor five times using the same target with different random seeds and the high precision BCUT descriptors. A single target was randomly selected (Figure 6) from among the 67 targets for which EA-Inventor failed to find a "0-Score" structure in the original Study 24. The results of these five EA-Inventor runs, shown in Table 4, demonstrate that this target can, indeed, be found using multiple EA-Inventor runs, as expected. It is also worth noting that in these cases, as for all cases in Studies 24 and 25, BCUT descriptors yielded only the "true positive" and no "false positive" results.

Precision. As discussed in the Introduction, the primary purpose of this study was to determine the extent to which compound descriptors can be shared for constructive purposes without revealing the identity of the compounds. We have suggested[17] that one way to achieve this objective is to provide the descriptors with sufficient precision for QSAR/QSPR model development but with insufficient precision to enable software like EA-Inventor to deduce the identity of the compounds for which the measured data was provided.

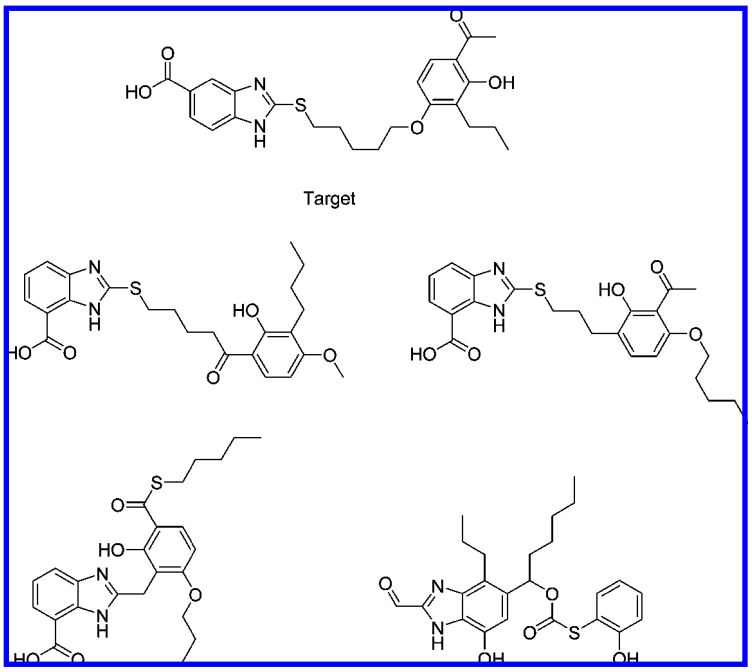

Figure 3. A target structure and four examples of "0-score" structures identified by EA-Inventor in Study 23. All 0-score structures have the same MlogP, ClogP, MW, N_{Hb-D}, N_{Hb-A}, N_{rot}, and N_{SCA} descriptor values as the target structure.

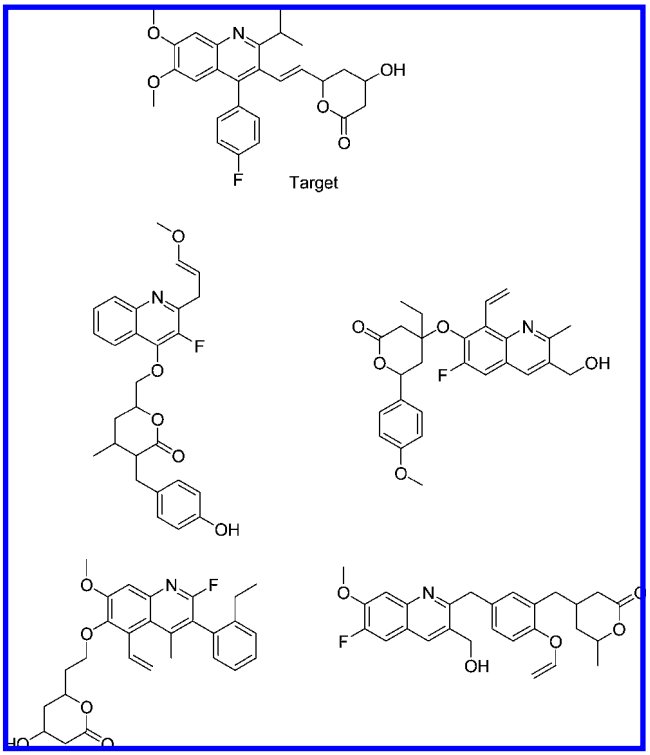

Figure 4. A second target structure and four examples of "0-score" structures identified by EA-Inventor in Study 23. All 0-score structures have the same MlogP, ClogP, MW, N_{Hb-D}, N_{Hb-A}, N_{rot}, and N_{SCA} descriptor values as the target structure.

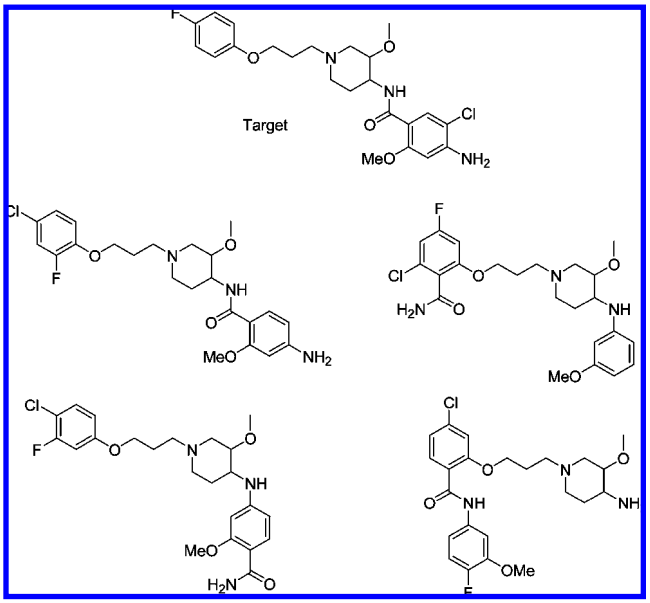

Figure 5. A target structure and four examples of "0-score" structures identified by EA-Inventor in Study 28. All 0-score structures have the same MACCs-vector descriptor values as the target structure.

For example, whether the MW value used for model development is 365.3 or 365.297 is of no real consequence even though there are *far* fewer compounds with MW values of 365.297 ± 0.001 than there are compounds with MW values of 365.3 ± 0.1.

In order to explore the notion that reduced precision prevents the deduction of chemical structure from descriptors, various levels of precision were selected for study. We observed that it was not possible to deduce chemical structure from the individual descriptors we studied, even at the highest precision possible. However, Studies 24, 25, and 26 (using a combination of BCUT descriptors provided with decreasing precision) clearly show that reducing the precision to three significant figures significantly reduces the ability to uniquely deduce the target structure.

The results from Study 27 are quite remarkable. In this study, a combination of Lipinski-like descriptors (Study 23) and low-precision BCUT descriptors (Study 26) was examined. While neither the combination of Lipinski-like descriptors (MlogP, ClogP, MW, N_{Hb-D}, N_{Hb-A}, N_{rot}, and N_{SCA}) alone nor the low-precision BCUTs alone were capable of

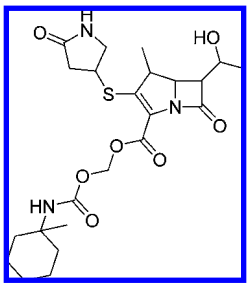

Figure 6. The target structure used for Studies 24_a, 24_b, 24_c, 24_d, and 24_e.

Table 4. Results of Repeated Attempts To Deduce the Structure of a Single Target Structure

study no.	descriptor[a]	0-score found?	structures with 0-Score[b]	invented equals target
24_a	(1−5) 2D-BCUT (7)	N	0	N
24_b	(1−5) 2D-BCUT (7)	Y	1	Y
24_c	(1−5) 2D-BCUT (7)	N	0	N
24_d	(1−5) 2D-BCUT (7)	N	0	N
24_e	(1−5) 2D-BCUT (7)	Y	1	Y

[a] The number in parentheses after the descriptor name is the "precision" (the number of significant figures after the decimal). [b] The number of structures that EA-Inventor was able to find with the same descriptor value as the target molecule. The maximum number of "0-scores" possible is 100.

uniquely revealing chemical structure, combining these two types of descriptors significantly improves the chances of uniquely determining the structure of the target. This demonstrates that caution should be used when combining descriptors. Combinations of "safe" descriptors may not be safe.

Guidelines for Safe Exchange of Data. As we have demonstrated, the extent to which structure can be deduced from descriptors depends, in part, upon the precision with which the descriptors are provided. Given the magnitudes of the inevitable errors in measured activities or properties, the statistical significance of QSARs or QSPRs is limited by the precision of the measured values, not by the precision of the descriptors. Thus, we can share descriptors without sharing structures simply by imposing sensible limits on the precision with which the corresponding descriptor values are communicated.

From these studies, we have developed the following general guidelines for "Safe Data Exchange", with the important caveats that combinations of too many "safe" descriptors may not be safe. We recommend that (1) Studies similar to those reported here be performed with the particular set of descriptors of interest to ensure that the chemical structures cannot be deduced from the information that will be shared. (2) Integral feature-counts (e.g., NHb-d, NHb-a, Nrot, NHal, and NAro) be used "as is". They do not provide sufficient information to reveal structure but they *can* be useful for QSAR. (3) Fingerprints and the corresponding integer-vectors need not and should not be used—they can reveal analogs or isomers of target structure. (4) Real-valued descriptors can and should be purposefully "approximated" by lowering the precision or adding a random increment too small to compromise QSAR or QSPR models but large enough to thwart attempts to deduce structure.

CONCLUSION

Our studies show that, unless sufficient precautions are taken, EA-Inventor is able to derive a chemical structure or a set of closely related analogs from some commonly used descriptors. Based on the results of our studies, a set of guidelines or recommendations for safely sharing chemical information without revealing chemical structure has been proposed. We also describe a procedure for assessing the risks of revealing chemical structure when exchanging chemical descriptor information. The procedures presented are generic and can be applied to any chemical descriptor or combination of descriptors and to any set of structures to enable a decision about whether the exchange of information can be done without revealing the chemical structures.

ACKNOWLEDGMENT

The authors wish to thank Dr. Robert Clark for his helpful comments in preparing this manuscript.

REFERENCES AND NOTES

(1) ACS Symposium. Safe Exchange of Chemical Information: Can Relevant Chemical Information be Exchanged without Disclosing Chemical Structure; 229th ACS National Meeting − Spring 2005, San Diego, CA.
(2) *J. Comput.-Aided. Mol. Des.* **2005**, 19 (9/10) and articles therein.
(3) Tetko, I. V.; Abagyan, R.; Oprea, T. I. Surrogate data − a secure way to share corporate data. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 749−764.
(4) See for example, Weininger, D. Method and apparatus for designing molecules with desired properties by evolving successive populations. U.S. Patent 5,434,796, 1995.
(5) Available from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144.
(6) BioByte Corp., 201 W. Fourth Street, Suite 204, Claremont, CA 91711−4707.
(7) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 339−353.
(8) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(9) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.
(10) MDDR: MDL Drug Data Report, MDL Information Systems, 2000.
(11) The five BCUTS used were specified as follows: 0.247420 0.542068 NLScaling[0.393220,0.148893]:BCUT, diag = Gast.-Huckel charges, off-diag = 10 * Burden's nums * 0.100, remove Hs, high eig; 2.62315 5.22006 NLScaling[3.99781,1.63352]:BCUT, diag = h-bond acceptor, off-diag = 10 * Burden's nums * 0.900, remove Hs, high eig; 1.67310 2.64211 NLScaling[2.16149,0.499074]:BCUT, diag = h-bond donor, off-diag = 10 * Burden's nums * 0.600, remove Hs, high eig; 2.63573 3.60850 NLScaling[3.14248,0.517471]:BCUT, diag = tab polar (A3), off-diag = 10 * Burden's nums * 0.500, remove Hs, high eig; and -1.54624 -0.726248 NLScaling[-1.15932,0.438683]:BCUT, diag = tab polar (A3), off-diag = 10 * Burden's nums * 0.750, remove Hs, low eig.
(12) MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577.
(13) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−80.
(14) See for example: Bologa, C.; Allu, T. K.; Olah, M.; Kappler, M. A.; Oprea, T. I. Descriptor collision and confusion: Toward the design of descriptors to mask chemical structures. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 625−635.
(15) Hicks, M. G.; Jochum, C. Substructure search systems. 1. Performance comparison of the MACCS, DARC, HTSS, CAS Registry, MVSSS, and S4 substructural search systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191−199.
(16) Willett, P. A Review of Chemical Structure Retreival Systems. *J. Chemom.* **1987**, *1*, 139−155.
(17) Shen, L.; Smith, K. M.; Masek, B. B.; Pearlman, R. S. Sharing Chemical Information without sharing Chemical Structure; Safe Exchange of Chemical Information Symposium, 229th American Chemical Society National Meeting, Spring - 2005.

CI600383V