Journal of Computer-Aided Molecular Design (2005) 19: 625–635
DOI 10.1007/s10822-005-9020-4

Descriptor collision and confusion: Toward the design of descriptors to mask chemical structures

Cristian Bologa[a], Tharun Kumar Allu[a], Marius Olah[a], Michael A. Kappler[b] & Tudor I. Oprea[a,c,*]
[a]*Division of Biocomputing, University of New Mexico School of Medicine, MSC11 6145, Albuquerque, NM, 87131, USA;* [b]*Daylight Chemical Information Systems, Inc., 441 Greg Avenue, Santa Fe, NM, 87501, USA;* [c]*Sunset Molecular Discovery LLC, 1704 B Llano St., Suite 140, Santa Fe, NM, 87505, USA*

Received 10 August 2005; accepted 6 October 2005

Key words: chemical fingerprints, ChemNavigator, descriptor collision, descriptor confusion, masking chemical structures, PLS, QSAR, SMILES, WOMBAT

Summary

We examined "*descriptor collision*" for several chemical fingerprint systems (MDL 320, Daylight, SMDL), and for a 2D-based descriptor set. For large databases (ChemNavigator and WOMBAT), the smallest collision rate remains around 5%. We systematically increase the "descriptor collision" rate (here termed "descriptor confusion"), in order to design a set of "descriptors to mask chemical structures", DMCS. If effective, a DMCS system would not allow third parties to determine the original chemical structures used to derive the DMCS set (i.e., reverse engineering). Using SMDL keys, the "confusion" rate is increased to 45.6% by eliminating those keys that have a low frequency of occurrence in WOMBAT structures. We applied an automated PLS engine, WB-PLS [Olah et al., J. Comput. Aided Mol. Des., 18 (2004) 437], to 1277 series of structures from 948 targets in WOMBAT, in order to validate the biological relevance of the SMDL descriptors as a potential DMCS set. The "reduced set" of SMDL descriptors has a small loss of modeling power (around 20%) compared to the initial descriptor set, while the collision rate is significantly increased. These results indicate that the development of an effective DMCS is possible. If well documented, DMCS systems would encourage private sector data release (e.g., related to water solubility) and directly benefit public sector science.

Abbreviations: CMR – calculated molecular refractivity; ClogP – program produced by BioByte Corp., Claremont, CA; Daylight/DY – Daylight Chemical Information Systems; DMCS – descriptors to mask chemical structures; DMSO – Dimethylsulfoxide; DPISMR – the NIH Small Molecule Repository as organized by DPI; LogP – the logarithm of the octanol-water partition coefficient; $LogS_w$ – the logarithm of the (molar) aqueous solubility; MACCS – Molecular ACCess System, an MDL product; MDL – Molecular Design Limited; MLI – Molecular Libraries and Imaging initiative; NIH – National Institutes of Health; PLS – Partial Least Squares/Projection Latent Structures; QSAR – quantitative structure–activity relationships; SMDL – Sunset Molecular Discovery, LLC; SMILES – Simplified Molecular Input Line Entry Specification; WOMBAT/WB – WOrld of Molecular BioAcTivity database.

*To whom correspondence should be addressed. Tel.: +1-505-272-6950; Fax: +1-505-272-0238; E-mail: TOprea@salud.unm.edu

Introduction

The Molecular Libraries and Imaging initiative (MLI) at the National Institutes of Health (NIH) [1] will increase the availability of small molecules as chemical probes for basic research, via the NIH Small Molecule Repository (search for "DPISMR" in PubChem [2]). Focused on the early stages of drug discovery and chemical biology, which encompass target identification, assay development, biomolecular screening and hit-to-probe analysis [1], the MLI aims to bridge the cultural divide between the public and private sectors. Envisioned as "public sector science", these MLI activities could then be followed by industrial and academic efforts in lead identification and optimization, followed by clinical trials. Crucial to the DPISMR collection assembly, which is to collect a representative set of small molecules for biomolecular screening, is quality assurance with respect to, e.g., certain physico-chemical properties such as DMSO and water solubility – both deemed important from a practical standpoint. Since measuring up-front a significantly large number of molecules (prior to selection into DPISMR) is not a feasible task (physical samples are required), one can employ computational methods to address this issue. However, there is a general consensus that existing computational models for water and DMSO solubility prediction could be improved. This, however, requires data from both the private and public sector to be released in such a manner that predictive models can be derived and become available.

Thus, one particular aspect of cheminformatics [3] has become relevant in the context of the MLI initiative: The ability to exchange *meaningful chemical information*, and to integrate that chemical information into knowledge. This is an explicit goal in the attempt to exchange chemical information and has been contemplated by scientists from both the industrial and in particular the academic sector. A considerable amount of experimental data such as solubility, melting points, $LogD_{7.4}$, cytochrome P450 inhibition/substrates, metabolic stability, toxicity – *in vitro* and in animals, human data (clinical pharmacokinetic for failed and successful trials) are available in the private sector. Such data are never made available for general use. Some scientists from the pharmaceutical industry are willing to share information with their academic colleagues, but cannot do so for proprietary and legal considerations.

However, there is a continuing discussion related to the safe exchange of chemical information. It is generally believed that one can reverse engineer chemical structures from descriptors [4]. Although such claims were not documented in peer-reviewed literature, one can imagine the use of a problem-solving algorithm such as a genetic algorithm [5] in conjunction with a large database, e.g., ChemNavigator's iResearch Library™ [6] or Beilstein [7], where appropriate descriptors are computed; with this procedure, one could conceivably converge to a structure that – given the descriptor space – is remarkably similar, or perhaps identical, to the "target structure" (i.e., match or come close to the unknown structure's fingerprint) with relative ease [8]. Such claims are now documented in respect to, e.g., the *Signature* descriptor system [9] and lead to the overall conclusion that it is not safe to release chemical information given the risk that structures might be reverse-engineered, in particular since the party releasing such information desires to keep its chemical structures secret.

To safely share chemical information, we need an uncrackable system that masks chemical structures. Useful information could then be provided to the public sector, which could then develop novel tools and validate models that benefits both public and private sectors of science, such as the DPISMR effort within the MLI initiative. Although, on the surface, some may question the benefits to the industry, the masking of chemical structures allows them to tap into the academic expertise without compromising intellectual property.

In this paper we discuss the aspect of "*descriptor collision*", where different chemical structures have the same identical set of descriptors – which in turn points to the possibility that safe exchange of chemical information is possible, if one deliberately develops a system where multiple collisions (or confusion) occurs. We distinguish "descriptor collision", where just by mere chance different chemical structures are mapped onto the same descriptor set, from "*descriptor confusion*", where the effort to increase descriptor collision is done by design. We highlight cases of "descriptor collision"; then outline how one could build such a "reduced set" of "confused descriptors" – generically termed "descriptors to mask chemical

structures'', DMCS, which remains relevant to quantitative structure–activity relationships (QSAR); finally we validate such a ''reduced set'' of DMCS on over 1000 QSAR series.

Materials and methods

Databases

To establish the reliability of fingerprint technology in mapping structures, we investigated structures from WOMBAT [10, 11] and the iResearch Library™ [6] with several descriptor systems. WOMBAT is a database for medicinal chemistry [10] that stores chemical structures and associated experimental biological activity data. Information about the original bibliographic reference from where the data was indexed is recorded, together with several calculated and experimental properties (e.g., bioactivities) for each structure. The iResearch Library™ from ChemNavigator is a database that caters to pharmaceutical research; it collects virtual (yet feasible) and existing commercial compounds from more than 200 suppliers.

For each database, we extracted all the unique, non-stereoisomeric (i.e., no R/S or E/Z isomers) SMILES strings [12], and found 98,785 structures in WOMBAT 2005.1 version [11], and 13,334,014 structures in iResearch Library™ 01/2005 version, respectively. Different types of descriptors (*vide - infra*) were computed for these two databases, and each pair of descriptors for different compounds were compared for identity. To further understand the ability of fingerprints to describe scaffolds, we took the WOMBAT 2005.1 SMILES strings and suppressed atom and bond types; this resulted in 70,211 unique graphs. This allows us to compare the number of unique fingerprints for unique structures, as well as the number of unique fingerprints for unique graphs.

Descriptors and fingerprints

Two-dimensional (2D) descriptors, i.e., descriptors that do not use information related to the three-dimensional (3D) characteristics of model compounds, were used in this study to evaluate the WOMBAT database. The 2D descriptors are *size-related: e.g.* molecular weight, calculated molecular refractivity – CMR [13], etc.; *hydrophobicity-related:* the logarithm of the octanol–water partition coefficient, LogP [14]; we used ClogP [15], and the logarithm of the (molar) aqueous solubility [16, 17] ($LogS_w$) as implemented in ALOGPS [18]; descriptors related to *electronic effects:* CMR, the (tabulated) estimated polarizability [19], partial atomic charges based on the Gasteiger–Marsili method [20], etc.; *hydrogen bonding descriptors* that include counts [21] of hydrogen bond acceptors or donors; and *topological indices* [22] derived from connectivity [23] matrices [24].

Binary representations for chemical structures characterization, also called *molecular* (or chemical) *fingerprints*, capture different molecular descriptors like atomic distances, pharmacophore patterns or unique structural paths. Two basic types of fingerprints are distinguished: *structural keys* and *hashed fingerprints*. The individual bits in structural keys (also termed ''keys'' or ''keybits'') are set to 1 or 0 depending on the presence or the absence of a particular chemical fragment from a predefined dictionary of fragments. As a prototype for fingerprint comparison, we used the MDL fingerprints [25] generated with the Mesa Analytics and Computing *Fingerprint* module, which reads SMILES strings as input and creates 320 bit structural MDL-MACCS key representations [26]. For hashed fingerprints, no fragment dictionary is used, thus there is no direct relationship between bits and features. We used the Daylight (DY) fingerprints, a path-based approach from Daylight Chemical Information Systems [27]. DY fingerprints perform a subgraph ''identification'': Starting from each atom and extending down each bond (including bond type), the entire molecule is assigned a unique pattern described by a small number of positions (e.g., 4 or 5) along the fixed-length fingerprint. Chain length is assigned a maximum length (typically 8). Although typical sizes for DY fingerprints are 1024 or 2048 bits in length, we used several other bit-lengths, ranging from 32 to 32,768 to monitor the number of unique fingerprints. When comparing databases, we used the 2D descriptors, the MDL 320 keys and the DY 512, 1024, and 2048 fingerprints.

To increase ''descriptor collision'', we decided to work with the SMDL (Sunset Molecular Discovery LLC [11]) fingeprints. As described earlier [28], these were inspired by the MDL 320 keys [25] and the CATS (chemically advanced template search) concept [29]. Our initial set of SMDL

keys, aimed at capturing topologically-relevant pharmacophore substructures, was validated [28] against 1632 QSAR series with at least 25 compounds per series extracted from WOMBAT 2004.1. For this study, the SMDL keys were reduced to the unique SMDL subset (from 504 initially to 403) where all non-carbon atoms are collectively defined as non-carbon. In other words, atoms such as N, O, S, P are no longer assigned their specific element, but are collapsed into the [!#6] SMARTS [30] atom type. For example, using this setting, an ether oxygen and a secondary amine are identified by the same fingerprint. The SMARTS definitions were further optimized using SMACK [31]. All SMDL keys were matched as SMARTS in WOMBAT and ChemNavigator using an in-house routine based on OEChem [31].

Relevance testing for DMCS descriptors

To generate a "reduced set" of DMCS, and to test for biological relevance, we used the SMDL keys. To increase the level of "descriptor confusion", different sets of SMDL keys were selected in two ways: (a) we look at those SMDL keys that are present in at least 10,000, 20,000 or 30,000 canonical isomeric SMILES from WOMBAT 2005.1 – as *binary* fingerprints; and (b) we look at those SMDL keys that occur in at least 10,000, 20,000 or 30,000 instances for the same set of SMILES – as (fragment) *counts*. Compared to binary fingerprints, fragment counts have the advantage of introducing some degree of non-linearity to linear modeling in QSAR. For the QSAR study, duplicates are reported with respect to the entire WOMBAT dataset (i.e., 98,785 canonical non-isomeric SMILES), since the QSAR modeling process used all available information related to structural input, including chirality.

Biological validation was performed using the PLS (Partial Least Squares [32]) engine from Sunset Molecular Discovery, WB-PLS [28]. Using WB-PLS, QSAR series were automatically extracted from the WOMBAT 2005.1 database. Two constraints were applied: (a) at least 25 structures with biological activity records in one QSAR series; (b) one activity column (Y) present for all molecules in each QSAR series [28]. For series with multiple activities, one QSAR series was generated for each activity block with respect to the first condition. An additional constraint, the ΔY criterion, discussed previously [28], was introduced in this study: All selected QSAR series span at least 2 log units for the dependent variable Y, with at least 20% of the compounds within one log-unit interval. Thus, only 1277 series compared to 1632 in the previous study [28] satisfy these criteria, even though the number of unique structures has increased from 64,539 (2004.1) to 104,230 (2005.1) when using DY canonical isomeric SMILES.

These series were used as input to the WB-PLS engine to test the explanatory and predictive power of "confused descriptors". For all PLS models discussed here, cross-validation was performed by randomly dividing each series into seven groups of similar size, then setting aside 1 out of 7 groups (CV7 cross-validation). The cross-validation statistic parameter (q^2) is reported here; q^2 is the cumulative fraction of the variance in the set-aside Y's that can be predicted by the PLS model derived on the remainder (6/7) of the dataset.

Results and discussion

Descriptor collision

Two aspects of data handling related to the safe exchange of chemical information are analyzed: "*descriptor collision*", i.e., the situation where non-identical chemical structures have the same identical set of descriptors by chance, and "*descriptor confusion*" where multiple non-identical structures have the same identical set of descriptors by design. *"Descriptor collision"* was observed in both databases, as illustrated in Figures 1–3 and summarized in Table 1. The figures show that such collisions occur for all tested systems, and that multiple collisions may occur simultaneously. For example, a set of different structures that have the same descriptor values for MDL 320 keys, 2D descriptors and DY 2048 fingerprints is given in Figure 3. As summarized in Table 1, the MDL 320 keys and the DY-512 keys have similar collision rates ($\sim$15%) for the large set of chemicals offered by suppliers (ChemNavigator database). However, the MDL keys show collision rates comparable to the DY-2048 fingerprints (around 5%) when tested against the medicinal chemistry compounds in WOMBAT. These results confirm that the MDL 320 keyset is optimized for mining pharmaceutical databases [25] and that MDL 320

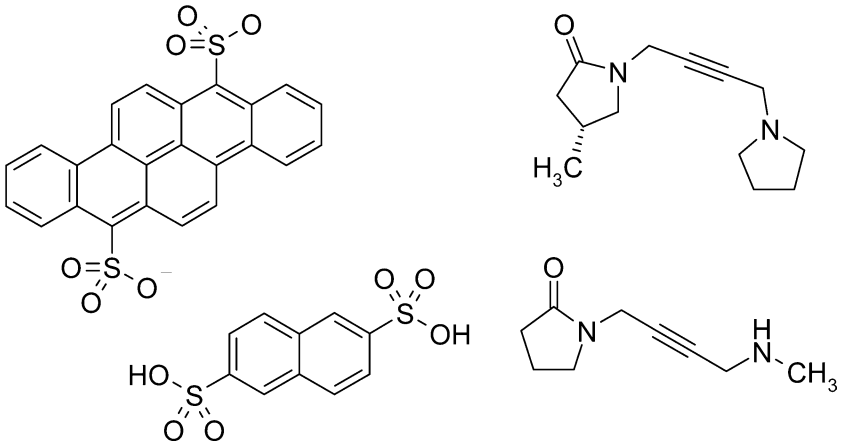

Figure 1. Examples of molecules with identical DY 2048 fingerprints.

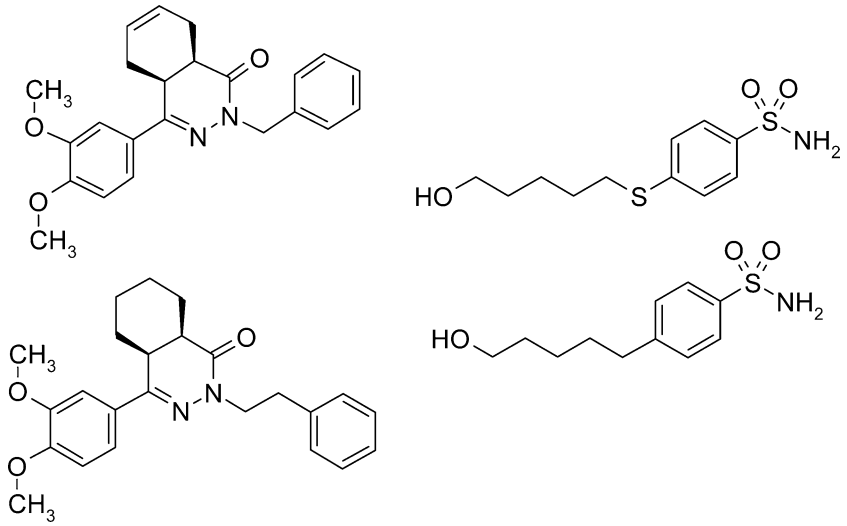

Figure 2. Examples of molecules with identical MDL 320 fingerprints.

and DY 2048 bitsets have comparable performance on such data. However, the DY 2048 set gives better results (lower collision rate) compared to the other 3 descriptor sets when used against the much larger set of supplier compounds.

The number of descriptor collisions decreases with the increase of fingerprint bit-size, as shown in Table 2. The WOMBAT 2005.1 database has 98,785 unique non-isomeric structures and 70,211 unique graphs (where information about atom and bond types is suppressed, i.e., 'alkanes'). The upper limit of 94,148 unique fingerprints for the DY-32,768 bits leads to 4427 duplicates (i.e., ~5%). The asymptotic increase in the number of unique fingerprints as the bit-size doubles indicates that some collisions are inherent to the system. However, the number of unique fingerprints for unique graphs peaks at a relatively small fingerprint size (DY-512), and is significantly below the number of initial unique graphs in the analyzed set. Since DY fingerprints are folded to decrease the search time for large databases, the results in Table 2 indicate that, in the absence of diverse atom and bond types, the number of up-to-eight paths generated by DY fingerprints is not an appropriate tool to describe alkanes – as noted elsewhere [33]. The results from Tables 1 and 2 show that, for all descriptor systems investigated here, it is not always possible to provide a 1:1 map between chemical structure and descriptor characterization.

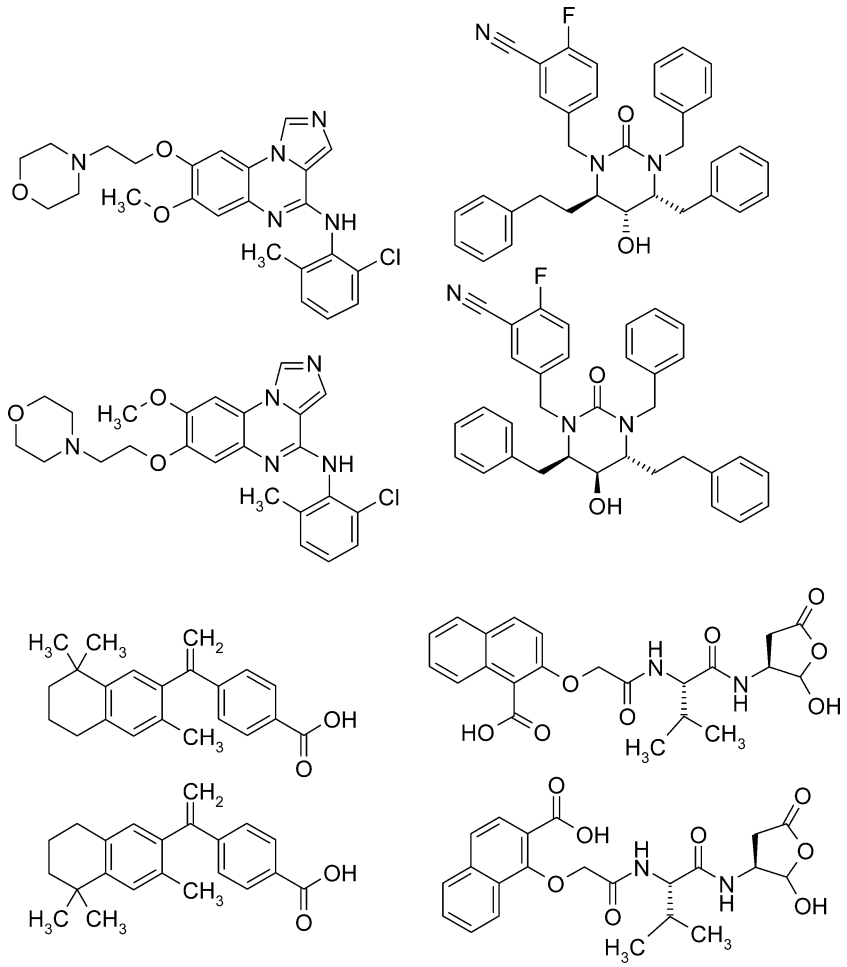

Figure 3. Examples of molecules with identical MDL320 keys, DY 2048 fingerprints, and 2D-properties.

Table 1. Descriptor collision rates for WOMBAT 2005.1 and the iResearch Library™.

Fingerprint type/size	WOMBAT duplicates[a]		iResearch Library™ duplicates	
	Number	%	Number	%
MDL – 320	4526	4.7	1,943,712	14.5
DY – 512	6684	6.8	2,029,981	15.2
DY – 1024	5496	5.6	974,395	7.3
DY – 2048	4979	5.1	702,333	5.3
SMDL – 403[b]	26,941	27.33	8,911,574	66.83

All non-unique patterns were considered duplicates. See text for further details.
[a]For WOMBAT we also computed descriptor collisions using the 2D descriptors reported elsewhere, [28] and found 433 (0.5%) duplicates.
[b]SMDL fingerprints were used to evaluate the number of duplicates; these fingerprints were deliberately engineered to increase "descriptor confusion".

Descriptor confusion

The above findings, in particular the collisions for MDL 320 keys and DY-512 fingerprints in ChemNavigator, indicate that, at least in theory, one could deliberately devise an information-rich, yet "confused" descriptor system, i.e., a chemical information exchange tool where chemical

Table 2. Fingerprint collisions for the WOMBAT 2005.1 dataset.

DY fingerprint size	Number of unique DY patterns	% Duplicates	Number of unique DY patterns for reduced graphs
32,768	94,148	4.5	178
16,384	94,123	4.5	178
8,192	94,087	4.6	178
4096	93,801	4.8	178
2048	93,596	5.1	178
1024	93,079	5.6	178
512	91,891	6.8	178
256	88,231	10.5	126
128	69,183	29.8	108
64	12,939	86.9	74
32	776	99.2	43

structure ambiguity is encouraged. The last row in Table 1 illustrates this concept, as the number of multiple non-identical structures having the same pattern (i.e., descriptor set) is deliberately increased compared to MDL and DY fingerprints. In particular, the initial set of 403 SMDL fingerprints were engineered to have a high collision rate, as illustrated for both the WOMBAT 2005.1 and ChemNavigator datasets. The collision rate in WOMBAT increases to over 27%; thus, almost 27,000 unique SMILES have at least one collision (many are also possible). This percentage is dramatically increased in ChemNavigator, to over 66.8% (almost 9 million structures). Because ChemNavigator contains large collections of combinatorial libraries that are currently available commercially, in contrast to (smaller) series of compounds that span over a decade of medicinal chemistry in WOMBAT, there appears to be a higher degree of degeneracy in ChemNavigator, compared to WOMBAT.

We investigated the effect of further reducing the precision of SMDL keys in WOMBAT by increasing the confusion rate, i.e., by eliminating those keys that do not occur in at least 10,000, 20,000 and 30,000 SMILES (binary) or instances (counts), respectively. The "*confused descriptors*", designed around a "reduced set" of SMDL keys, indicate that the conceptual development of an effective DMCS set (descriptors to mask chemical structures) is possible, as illustrated in Table 3 (column 3). Thus, the number of unique patterns drops from ~71,500 when using the 403 (binary) SMDL keys, to ~53,500 when using only a subset of 114 keys present in at least 30,000 SMILES. This trend was not reproduced when monitoring the number of unique patterns generated by fragment counts, as only an insignificant drop is observed when going from all 403 counts (~88,700 unique patterns) to 194 counts present in at least 30,000 instances (~88,600 unique patterns), respectively.

The "descriptor confusion" concept was further confirmed by an extensive analysis of the degeneracy rate in the WOMBAT 2005.1 and ChemNavigator – as summarized in Figures 4 and 5 and Table 4. Degeneracy indicates the number of unique canonical non-isomeric Daylight SMILES that have the same pattern when using the same SMDL keys as reported in Table 3. The degeneracy rate, illustrated in Figures 4 and 5, indicates how the number of unique patterns is influenced by the descriptor system, i.e., how "descriptor confusion" influences uniqueness. In both the WOMBAT (Table 3) and ChemNavigator (Table 4), the degeneracy rate remains quasi-constant for fragment counts, but shows a significant increase for binary fingerprints. For example, there are 34,513 SMILES in WOMBAT and 4,758,843 SMILES in ChemNavigator that have a degeneracy between 2 and 10 when using all 403 fingerprints. This number increases to 46,601 in WOMBAT (see also Figure 4), and decreases to 3,531,380 when using the "confused" set of 114 fingerprints (see also Figure 5), for the same degeneracy rate. The degeneracy rate between 11 and 100 further increases in the ChemNavigator dataset, from 3,854,403 (when using 403 fingerprints) to 4,195,458 when using the "confused" fingerprints (see Figure 5). While the degeneracy

Table 3. Summary of QSAR validation results.

Binary descriptor set	Unique patterns		QSARs with $q^2 \geq 0.3$		Counts descriptor set	Unique patterns		QSARs with $q^2 \geq 0.3$	
	Total	Percent duplicates	Total	Percent		Total	Percent duplicates	Total	Percent
All FPs (403)	71,634	27.33	283	22.16	All CTs (403)	88,768	9.94	403	31.55
> 10KFP (232)	68,824	30.18	278	21.76	> 10KCT (250)	88,759	9.95	400	31.32
> 20KFP (170)	62,667	36.43	260	20.36	> 20KCT (219)	88,647	10.06	390	30.54
> 30KFP (114)	53,563	45.66	236	18.48	> 30KCT (194)	88,550	10.07	394	30.85

The number of unique patterns reflects the perception of uniqueness by each descriptor system, not the number of unique chemical structures.

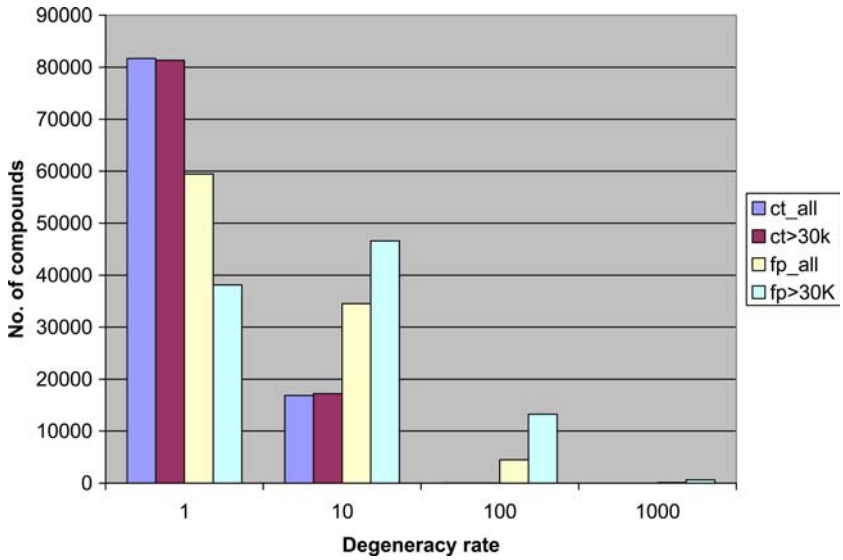

Figure 4. Degeneracy rate for Wombat collection.

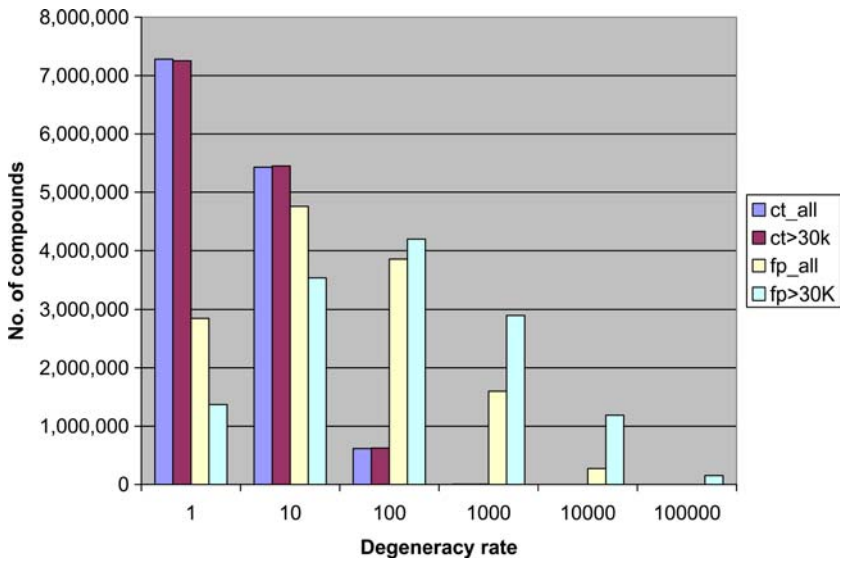

Figure 5. Degeneracy rate for ChemNavigator collection.

Table 4. Unique patterns for the ChemNavigator database, using the SMDL descriptors validated on WOMBAT activities (see also Table 3).

Binary descriptor set	Unique patterns		Counts descriptor set	Unique patterns	
	Total	Percent duplicates		Total	Percent duplicates
All FPs (403)	4,422,440	66.83	All CTs (403)	9,277,035	30.42
> 10KFP (232)	4,030,158	69.76	> 10KCT (250)	9,276,670	30.43
> 20KFP (170)	3,407,816	74.44	> 20KCT (219)	9,252,845	30.61
> 30KFP (114)	2,529,849	81.03	> 30KCT (194)	9,252,792	30.61

rate does not change for fragment counts, we can state that it increases when reducing the descriptor set for fingerprints from 403 to 114 – as observed for both databases. Another way to look at this type of data is to evaluate the average degeneracy per structure, after excluding the number of compounds with unique patterns. In other words, we wanted to monitor degeneracy rate after eliminating structures that map onto unique patterns, and find out if there is a net change. After eliminating unique structures/patterns, the average degeneracy increases from 3.21 (WOMBAT) and 6.66 (ChemNavigator) for the 403 SMDL binary fingerprints, to 3.91 (WOMBAT) and 10.31 (ChemNavigator) for the 114 fingerprints. These pairs of numbers reflect the net increase in degenerate states for those structures that are no longer mapped to a unique set of descriptors; since this is a net increase, we can further conclude that the "descriptor confusion" effort is effective in masking chemical structures.

Relevance testing for DMCS descriptors

The testing for relevance as DMCS descriptors was done by evaluating modeling performance, using CV7 q^2 cross-validation values, i.e. testing the performance of these systems to generate QSAR equations for a diverse set of biological targets and chemical structures. Overall, the combined dataset consisted of 1277 series derived from 948 unique structure–activity series, or targets (some had multiple Y columns), totaling 50,925 activity–structure pairs. Since relevance testing is performed on a considerable number of series, the distribution of q^2 values (Figure 6) is likely to address the key question, namely are "reduced set" descriptors of any use as DMCS, given that "descriptor confusion" is designed to reduce the 1:1 mapping between structures and descriptors. In other words, *does the (intended) loss of information specificity influence the modeling ability of the descriptor set?*

As previously described [28], we have chosen a threshold of $q^2 \geq 0.3$ as a critical value based on the empirical observation that by eliminating between 1 and 5 objects from an initial set of 25, the q^2 value can increase to 0.5 or higher, i.e., the model achieves a (marginally) significant internal predictivity. Emphasis is given, in this discussion, to the proof-of-concept for DMCS development. Thus, aspects related to outlier detection and interpretation of QSARs, as well as those related to automated QSAR generation and processing, are neglected.

As illustrated earlier, the method of increasing descriptor confusion results in a significant reduction of the number of unique patterns, but has a lesser impact on the number of valid QSARs derived using those descriptors (see Table 3 and Figure 6). By considering the area-under-the-curve criterion, we note that the green dashed line (> 30KCT, 194 SMDL counts) and the indigo line (> 30KFP, 114 SMDL binary keys) do not give the best results (Figure 6) compared to the other descriptor sets. This is not surprising given

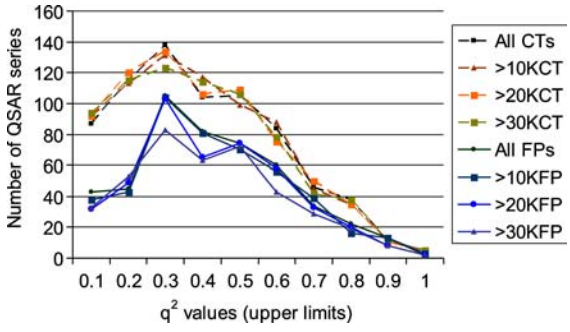

Figure 6. Distribution of cross-validated q^2 (using 0.1 intervals).

that significant QSARs are derived for only 236 series (83.4%) of the 283 series that were successfully modeled by the initial SMDL binary keyset, and for 394 series (97.8%) of the 403 series successfully modeled by the initial SMDL fragment counts set, respectively. This drop in relevance (see Table 3 for the overall percentage) is paralleled by a drop in unique patterns from ~71,500 to ~53,500 for the binary set, but not for the SMDL counts, respectively. The performance of the 114 SMDL binary keys is comparable, overall, to that of the MDL 320 (binary) keys (22.4% series with $q^2 \geq 0.3$ for CV7), whereas the 194 SMDL fragment counts behave comparably to the earlier set (504) of SMDL counts (31.1% series with $q^2 \geq 0.3$ for CV7), as tested on the 1632 series from WOMBAT that we studied previously under similar conditions [28]. Both descriptor sets appear to have superior performance compared to that of the 2D properties set (15.1% series with $q^2 \geq 0.3$ for CV7 for the 1632 series; not modeled on the current 1277 series).

The "confused descriptor set" of 114 SMDL binary keys has 53,563 unique patterns, when tested on the 98,785 unique WOMBAT SMILES. Thus, descriptor collision rate almost doubles, from ~27% in the SMDL-403 set to approximately ~45% in the SMDL-114 set, with ~20% drop in modeling ability of QSAR datasets. This result answers our key question in a favorable manner, showing that loss of information specificity is not necessarily coupled with a significant reduction in the ability of these descriptor sets to generate significant QSARs. It further supports our initial hypothesis, namely that the design a DMCS set with increased descriptor confusion is possible without significant loss of relevance for biologically related properties.

Conclusions

In this paper we document how SMDL keys, originally built around the topological pharmacophore concept, can be engineered to increase descriptor collision rates from ~5% (as illustrated by MDL-320 and DY-2048 keys) to ~45%, with only a limited loss of modeling power across large numbers of QSAR series. By combining such "reduced sets" of descriptor keys with, e.g., binned 2D descriptor values (e.g., for MW, ClogP, CMR, and topological indices), one can further extend descriptor relevance in QSAR models and observe acceptable (safe) "confusion" rates. An arbitrary proposal for safe exchange of chemical information [9] considers that, for a DMCS system to be "safe", less than 25% of the compounds should have degeneracy below 10 (i.e., mapping one pattern into at least 10 chemical structures). The DMCS proposed here does not yet meet this criterion for WOMBAT, but it certainly moves in this direction if one considers the results[1] from the ChemNavigator database. We conclude that this initial attempt to increase "descriptor confusion" while maintaining relevance for biological QSARs is successful. We believe that the design of an effective DMCS system is possible. Our attempts, focused strictly on therapeutically relevant biological targets (e.g., serine proteases, nuclear hormone receptors, class A G-protein coupled receptors, etc.), complement the effort of developing a safe exchange system for physico-chemical properties such as LogP [8], where actual disclosure of chemical structures may not be required.

As the MLI initiative plans to acquire and screen between one half to one million structures for biological activity, this requires a significant effort for *in silico* support, e.g., with respect to water & DMSO solubility prediction. Our report indicates that the development of an effective DMCS is possible. If well documented, this would encourage release of private sector data and result in a direct benefit to public sector science. A directed effort to derive DMCS in source code (open access) may facilitate the private sector release of important data. In an ideal scenario, tables of descriptors and measured properties would thus become available and multiple predictive models (e.g., using linear and non-linear methods, as illustrated here) could be derived. We anticipate that predictive models based on such DMCS systems will improve as large number of endpoints become available for

[1] When using SMDL-403 binary keys, we found 4589 (4.6%) WOMBAT structures and 5,727,197 (42.9%) ChemNavigator structures with degeneracy above 10. These values increased to 13,865 (14.1%) in WOMBAT and 8,281,658 (62.1%) in ChemNavigator respectively, when using the SMDL-114 binary keys.

model development, and will directly benefit public sector science such as the DPISMR compound acquisition effort associated with the MLI initiative.

Acknowledgments

We thank Jeremy (JJ) Yang from OpenEye Scientific Software (Santa Fe, NM) for advice on descriptor collision. This work was supported by New Mexico Tobacco Settlement Funds for Biocomputing (TKA, MO) and by the New Mexico Molecular Library Screening Center, NIH 1U54 MH074425-01 (CB, TIO).

References

1. Austin, C.P., Brady, L.S., Insel, T.R. and Collins, F.S., Science, 306 (2004) 1138.
2. The PubChem database is available online at the National Center for Biotechnology Information, http://pubchem.ncbi.nlm.nih.gov/ Last access on 21.10.05.
3. Hahn, M.M. and Green, R., Curr. Opin. Chem. Biol., 3 (1999) 379.
4. Filimonov, D. and Poroikov, V., J. Comput. Aided Mol. Des., 19 (2005) 10.1007/s10822-005-9014-2.
5. Weber, L., Curr. Opin. Chem. Biol., 2 (1998) 381.
6. The iResearch Library™ is available from ChemNavigator, Inc., http://chemnavigator.com/cnc/products/IRL.asp Last access on 21.10.05.
7. The Crossfire Beilstein database is available from Elsevier MDL, http://www.mdl.com/products/knowledge/crossfire_beilstein/index.jsp Last access on 21.10.05.
8. Tetko, I.V., Abagyan, R. and Oprea, T.I., J. Comput. Aided Mol. Des., 19 (2005) 10.1007/s10822-005-9013-3.
9. Faulon, J.L., Brown, W.M. and Martin, S., J. Comput. Aided Mol. Des., 19 (2005) 10.1007/s10822-005-9007-1.
10. Olah, M., Mracec, M., Ostopovici, L., Rad, R., Bora, A., Hadaruga, N., Olah, I., Banda, M., Simon, S., Mracec, M. and Oprea, T.I., In Oprea, T.I. (Ed), Chemoinformatics in Drug Discovery, Wiley-VCH, New York, 2005, pp. 223–239.
11. WOMBAT is available from Sunset Molecular Discovery LLC, http://www.sunsetmolecular.com/ Last access on 21.10.05.
12. Weininger, D., J. Chem. Inf. Comput. Sci., 28 (1988) 31.
13. Leo, A. and Weininger, D., CMR3. Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/, 1995.
14. Leo, A., Chem. Rev., 93 (1993) 1281.
15. Leo, A. and Weininger, D., CLOGP 4.0. Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/, 2001.
16. Ran, Y., Jain, N. and Yalkowsky, S.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1208.
17. Livingstone, D.J., Ford, M.G., Huuskonen, J.J. and Salt, D.W., J. Comput. Aided Mol. Des., 15 (2001) 741.
18. Tetko, I.V., Tanchuk, V.Y. and Villa, A.E., J. Chem. Inf. Comput. Sci., 41 (2001) 1407.
19. Glen, R.C., J. Comput. Aided Mol. Des., 8 (1994) 457.
20. Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.
21. Oprea, T.I., J. Comput. Aided Mol. Des., 14 (2000) 251.
22. Balaban, A.T., SAR QSAR Environ. Res., 8 (1998) 1.
23. Kier, L.B. and Hall, L.H. Molecular Connectivity in Structure–Activity Analysis. John Wiley, New York, 1986.
24. Basak, S.C., Balaban, A.T., Grunwald, G.D. and Gute, B.D., J. Chem. Inf. Comput. Sci., 40 (2000) 891.
25. Durant, J.L., Leland, B.A., Henry, D.R. and Nourse, J.G., J. Chem. Inf. Comput. Sci., 42 (2002) 1273.
26. MacCuish, J. and MacCuish, N., Measures software, Mesa Analytics and Computing LLC, Santa Fe, New Mexico, http://www.mesaac.com/ Last access on 21.10.05.
27. Daylight fingerprints are available from Daylight Chemical Information Systems, http://www.daylight.com/ Last access on 21.10.05.
28. Olah, M., Bologa, C. and Oprea, T.I., J. Comput. Aided Mol. Des., 18 (2004) 437.
29. Schneider, G., Neidhart, W., Giller, T. and Schmidt, G., Angew. Chem. Int. Ed., 38 (1999) 2894.
30. The SMARTS toolkit and SMARTS are available from Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/dayhtml/doc/theory.smarts.html; online SMARTS tutorial: http://www.daylight.com/dayhtml/doc/theory/smarts.html, 2005.
31. SMACK and OEChem are available from OpenEye Scientific Software, Santa Fe, New Mexico, http://www.eyesopen.com/products/applications/smack.html, 2005.
32. Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H., (Ed), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 523–550.
33. Kappler, M.A., Allu, T.K., Bologa, C. and Oprea, T.I., J. Chem. Inf. Model, 45 (2005) in preparation.