J Comput Aided Mol Des (2010) 24:521–551
DOI 10.1007/s10822-010-9346-4

Tautomerism in large databases

Markus Sitzmann · Wolf-Dietrich Ihlenfeldt ·
Marc C. Nicklaus

Received: 2 February 2010 / Accepted: 17 March 2010 / Published online: 29 May 2010

Abstract We have used the Chemical Structure DataBase (CSDB) of the NCI CADD Group, an aggregated collection of over 150 small-molecule databases totaling 103.5 million structure records, to conduct tautomerism analyses on one of the largest currently existing sets of real (i.e. not computer-generated) compounds. This analysis was carried out using calculable chemical structure identifiers developed by the NCI CADD Group, based on hash codes available in the chemoinformatics toolkit CACTVS and a newly developed scoring scheme to define a canonical tautomer for any encountered structure. CACTVS's tautomerism definition, a set of 21 transform rules expressed in SMIRKS line notation, was used, which takes a comprehensive stance as to the possible types of tautomeric interconversion included. Tautomerism was found to be possible for more than 2/3 of the unique structures in the CSDB. A total of 680 million tautomers were calculated from, and including, the original structure records. Tautomerism overlap within the same individual database (i.e. at least one other entry was present that was really only a different tautomeric representation of the same compound) was found at an average rate of 0.3% of the original structure records, with values as high as nearly 2% for some of the databases in CSDB. Projected onto the set of unique structures (by FICuS identifier), this still occurred in about 1.5% of the cases. Tautomeric overlap across all constituent databases in CSDB was found for nearly 10% of the records in the collection.

M. Sitzmann · M. C. Nicklaus (✉)
Chemical Biology Laboratory, Center for Cancer Research,
National Cancer Institute, National Institutes of Health, DHHS,
NCI-Frederick, 376 Boyles St., Frederick, MD 21702, USA
e-mail: mn1@helix.nih.gov

W.-D. Ihlenfeldt
Xemistry GmbH, Hainholzweg 11, 61462 Königstein, Germany

Keywords Database · Tautomers · Stereochemistry · Chemoinformatics · Small molecules

Introduction

Tautomerism is defined as the isomerization of a chemical compound according to the general scheme shown in Fig. 1 [1].

During this rearrangement, G is typically a single atom or group being transferred from X to Y. G acts as a leaving group during isomerization, X and Y serve as a donor and acceptor for G, respectively. Simultaneously with the transfer of G, pi electrons migrate in the opposite direction of G. Each of the atoms X, Y or Z can be of any of the atom types C, N, O, or S; G can be H, methyl, CH_2R, Br, NO, SR, or COR [2]. If it is a conjugated pi system, Y can be a larger group of atoms that allows the transfer of the pi electrons.

If the transferred group G is a proton the isomerization is called "prototropic tautomerism". This will be the only type of tautomerism discussed in this paper.

Irrespective of the type of rearrangement, any two different tautomers of the same chemical compound differ in their location of atoms and distribution of pi electrons. Therefore, different tautomers of a chemical compound may differ in their pattern of functional groups, double bonds, stereo centers, conformation, shape, surface or hydrogen-bonding pattern.

This difference can have an effect in any area of chemistry in which the static, classical connectivity between atoms is of importance, including computer-aided

$$G{-}X{-}Y{=}Z \rightleftharpoons X{=}Y{-}Z{-}G$$

Fig. 1 General isomerization scheme for tautomers

molecular design, property predictions, and chemoinformatics. Specifically, it may affect:

- calculation of physicochemical properties (pK_a, lipophilicity, solubility etc.)
- structure clustering and similarity searching (different fingerprints)
- database registration (identification of different tautomers of the same compound)
- virtual screening methods (different hydrogen-bonding patterns and H-donor or H-acceptor patterns)
- assignment to substance classes, e.g. may change the property "aromatic"
- predicted (or queried) reaction patterns which may be different for different tautomers [3–6].

A significant and often overlooked effect of tautomerism is that it can change the stereochemistry of a compound. There are two cases: The location of the double bond changes as illustrated in the scheme shown in Fig. 1, which might add or eliminate the presence of an E/Z stereo bond as shown in Fig. 2 (top). Likewise, migration of a double bond to a heretofore sp3 hybridized atom that was chiral removes this chirality, which, in a further tautomeric isomerization step, can be re-established, but with the opposite chirality—which effectively may result in the racemization of this stereo center (Fig. 2, bottom).

It is well-established that tautomerism (as a difference in representation *vs*. truly separable isomers) depends on conditions including pH, temperature and solvent [7]; it gains crucial importance in the above areas when it is likely to lead to interconversion under "normal" conditions, in which case different tautomeric representation would lead to an erroneous differentiation between isomers that are in reality the same compound ("stuff in the bottle"). It is typically not possible to probe the tautomeric situation experimentally for databases of any significant size. Likewise, while an experienced chemist is likely to be able to recognize cases of tautomerism when visually inspecting small sets of compounds, this approach is obviously not applicable to databases of thousands or even millions of structures. Computational tools are needed instead, and those need to be made tautomerism-aware. This has not always been the case in the past.

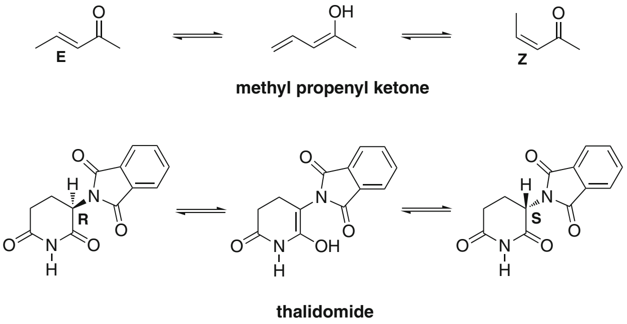


Fig. 2 Tautomerism can change stereochemistry. *Top*: change of E/Z geometry. *Bottom*: change of chirality

Specifically, when registering a new compound in a database, be it a vendor catalog, a commercial aggregation such as the *iResearch Library* (*iRL*) from ChemNavigator (Sigma–Aldrich) [8], or a public database such as PubChem [9], it is typically necessary to know if this compound is truly new, or may already be present but may have been previously entered as a different tautomer. This issue likewise comes up in the context of drug design, e.g. as the question: Are some of the members of my contemplated, *in silico*-designed, library perhaps present in a commercial catalog, but shown as a different tautomer? It has even been reported that tautomeric pairs were found in vendor catalogs for the same compound sold (unwittingly) as different products at different unit pricing [4, 10]. A less common but more drastic case may occur if all structures in an existing database are recalculated, possibly from a different raw source; which is what happened around 2000 with the NCI Database [11, 12] with the effect that, to our great surprise, about 100,000(!) out of ~250,000 structures appeared to have changed. Subsequent analysis revealed that in most of these cases, the tautomer represented had changed. It is thus not only satisfactory from a theoretical chemoinformatics point of view to correctly handle tautomerism issues, but also of significant practical importance.

The arguably next-best approach to experimentation, sufficiently high-level quantum-chemical computations, are not possible for large numbers of structures due to the enormous amount of computer resources required, not to speak of the difficulties of faithfully representing all environmental conditions in such runs. A rule-based chemoinformatics treatment is therefore usually the only practical approach. This implies that the outcome of such tautomer calculations is dependent on the exact rules used, whether they are implemented in a fixed way in the chemoinformatics tool used, modified by the user starting from a predefined rule set, or completely created "de novo" by the user. It is therefore to be expected that the results of our analyses would look quantitatively somewhat different if they had been conducted with different tools and thus tautomerism definitions. It would make little sense to say, "Your analysis must be wrong since we do find a different degree of duplication by tautomer overlap in our database than you report in your study."

It would be fascinating to try to compare computational tautomerism rules in an "experimental chemoinformatics" way: identify tautomer pairs (or n-tuples) among commercially available samples, based on different sets of tautomerism rules; purchase a number of such sample pairs; and test them by analytical chemistry methods such as NMR and mass spectrometry, possibly under systematically varied conditions (pH, temperature, solvent, etc.), to determine, at least statistically and based on that sample set, which rule sets better reproduce measured sample identity vs. sample difference. This can obviously not be done as part of this study since it is a large-scale project in its own right.

Our collection of small-molecule structures aggregated from numerous databases of very different origin, purpose, and size, has recently breached the 100 million record limit (see below). Its nature as one of the currently largest databases of *existing* small molecules, *vs*. very large databases of computer-generated structures [13], offers the unique opportunity to conduct all kinds of studies on a structure set that is not only highly statistically relevant by its sheer size but simply represents a good part of the real chemistry "out there."

We attempt to give some quantitative answers in this paper on how prevalent tautomer overlap (according to our tautomerism definitions) is in specific databases that make up our aggregated collection. One primary approach to this is to find a canonical representation independent on which tautomer was originally submitted. Depending on how such an approach is implemented, it does not preclude the possibility of keeping different original tautomeric forms for the registration in databases. We present an approach that achieves this based on our specifically crafted identifiers.

Methods

Data set

The data set used for this study was the aggregated database of structures collected by the Computer-Aided Drug Design (CADD) Group of the National Cancer Institute (NCI). This collection has been put together from a diverse set of small-molecule databases, and is referred to by us as the Chemical Structure DataBase (CSDB). It serves as the central small-molecule repository at the NCI CADD Group. It is a source of both commercially and otherwise available screening samples as well as of structural ideas in general for our internal CADD-type work, the basis for many of our public web services, and convenient fount of structures for chemoinformatics studies such as this. The current main sources for chemical structure records in CSDB are the ChemNavigator *iResearch Library* of commercially available screening samples [14] and PubChem [9]. Additionally, a few small-molecule database coming from other sources such government agencies and academic groups, as well as some vendor catalogs have been included. In its current version (Jan 2010), the CSDB indexes approximately 103.5 million original structure records, which represent about 70.6 million unique chemical structures [15].

It should be noted that CSDB comprises essentially all well-defined (external) databases used in the CADD Group's work, including proprietary and non-public ones. It is therefore a superset of the structure sets that are offered to the public in our various web-based tools and downloadable data sets [16]. The CSDB data set was used "as is" for this study (it is, after all, large enough); i.e. we did not specifically try to update all original databases in it, some of which are present in somewhat older versions. Our results can therefore not necessarily be taken as an assessment of the tautomeric situation of databases that are continuously being updated and/or curated and may be different in their current versions.

The PubChem data set was downloaded from the PubChem FTP site [17]. Since we generally need for our CADD work PubChem's assay results, too, which are only available in PubChem's Substance set (containing the original structures), we typically download this set. This gave us the structures from all of PubChem's sub-databases in their "rawest" form, i.e. the form least processed by PubChem relative to the representation submitted by the original provider. We did not additionally download the so-called Compound set, which contains the de-duplicated structures based on the normalization applied by PubChem.

Since we register structure records in CSDB that come from various sources and different chemical structure databases, a crucial step during the registration process is the normalization of the chemical structures. Normalization is needed because what is actually the same chemical may be encoded in different ways in different input databases if not the same database, be it due to certain chemical features of the structure that can lead to variable representation, for instance different tautomers or different resonance structures, or be it caused by ill-defined parts of the structure such as misdrawn functional groups, missing hydrogen atoms, missing charges or incorrect valences.

Structure normalization

Figure 3 illustrates the registration process for a new structure record to be entered in CSDB. This process has been entirely implemented on the basis of the chemical data management system CACTVS [18, 19].

CACTVS is able to read chemical structures from an extensive list of different file formats, which therefore

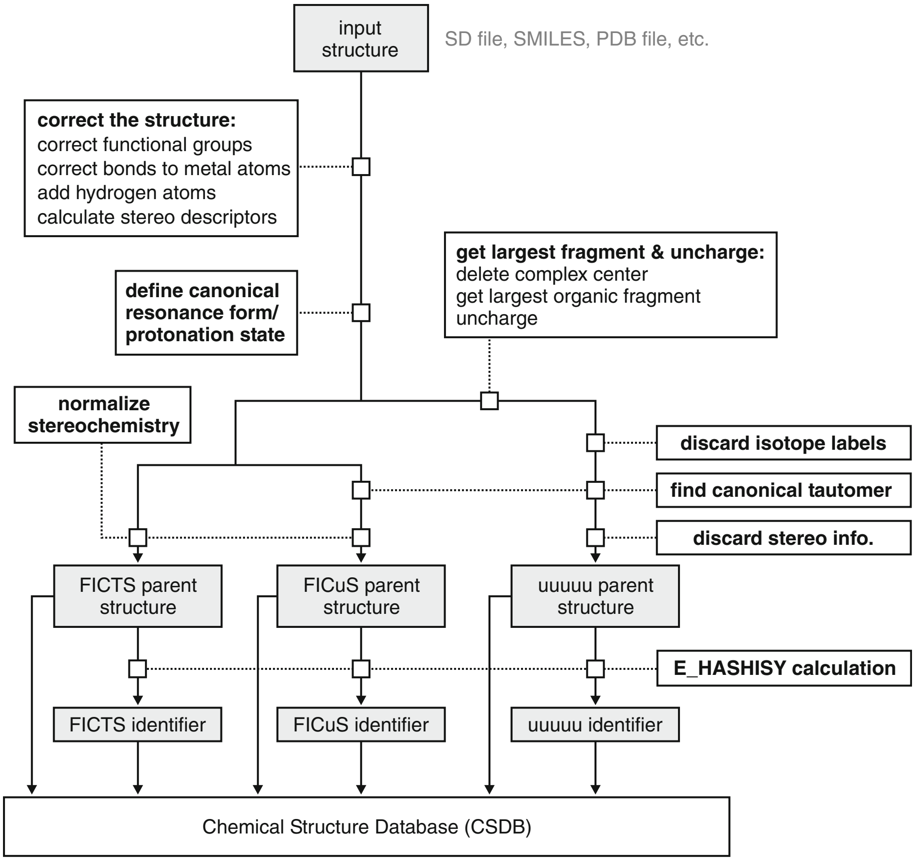


Fig. 3 Calculation of the NCI/CADD Chemical Structure Identifiers (FICTS, FICuS, uuuuu)

could in principle be all used in the registration of new compounds in CSDB. However, it has so far only been necessary to process databases using the SD file format for the addition of new records into CSDB.

As illustrated in Fig. 3, at the end of the registration process different parent structures are produced from the original input structure. The first step in this process are a few types of structure correction, which are meant to remove mostly error-based differences in the representation of what is really the same chemical. In contrast hereto, the several variants of our parent structures are obtained by removal of different subsets of chemical features that, if correctly assigned, represent truly different chemicals such as different stereoisomers, different salt forms, differently isotopically labeled compounds etc. This is described in more detail below.

For all these parent structures the corresponding NCI/CADD Chemical Structure Identifiers are generated [20]. For this, the E_HASHISY hash code function [21] in CACTVS is used, which calculates a 16-digit hexadecimal (64-bit unsigned) number from an arbitrary chemical structure. E_HASHISY represents a chemical structure very exactly as drawn, i.e. the hash code value changes as soon as connectivity, bond orders, atom types (including isotopes) or stereochemistry changes in the structure.

The parent structures obtained by the structure normalization process represent the original input structure with different levels of sensitivity to chemical features in the original structure. The sensitivity to specific chemical features is adjusted by switching on or off different algorithmic modules during the structure normalization process. Although we have implemented in total a set of eight variants of our identifiers, the most important ones are the "FICTS," "FICuS" and "uuuuu" parent structure and identifier, which are calculated for each structure registered in CSDB. The naming scheme behind these identifier designations has been explained elsewhere [20]. Briefly, the five letters "F," "I," "C," "T," and "S" stand for sensitivity to fragments, isotopic labeling, charges, tautomerism, and stereochemistry information, respectively, that may be present in the input structure. If, for a given identifier, sensitivity to any of these chemical features is switched off, the corresponding upper-case letter is replaced by a lower-case "u" (standing for "un-sensitive").

The FICTS parent structure and its identifier are thus a very close representation of the original input structure. The normalization procedure for both the parent structure and the identifier consists here mainly of a few corrections that fix and unify some typical drawing deficiencies and variations of how certain chemical features in chemical

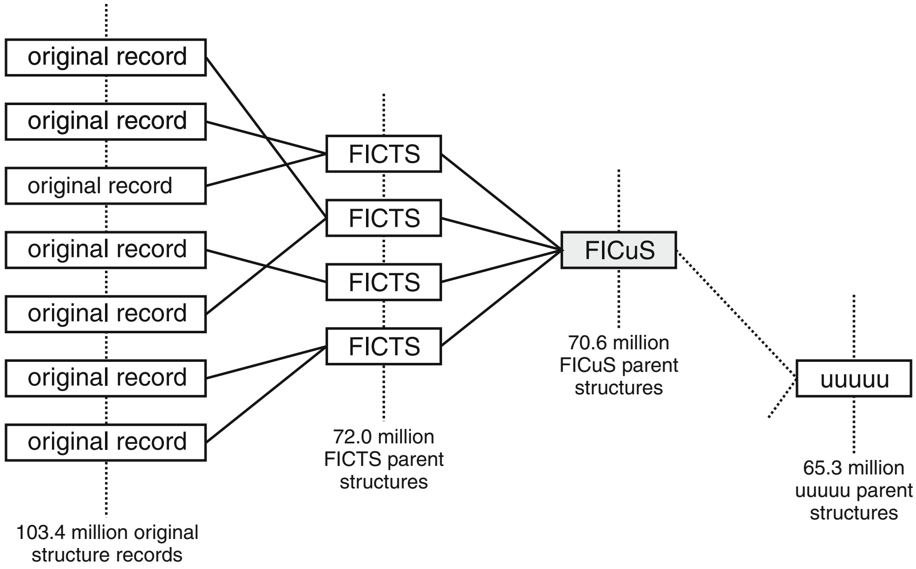

original record
original record
original record
original record
original record
original record
original record
103.4 million original structure records
FICTS
FICTS
FICTS
FICTS
72.0 million FICTS parent structures
FICuS
70.6 million FICuS parent structures
uuuuu
65.3 million uuuuu parent structures

Fig. 4 Relationship between the different NCI/CADD parent structures and identifiers after structure normalization

structures coming from different sources are specified (e.g. different drawing variants of functional groups). The FICTS representation of a structure is sensitive to fragments (such as counterions), isotopes, charges, and stereochemistry in the input structure as well as to the specific tautomer drawn.

The FICuS normalization procedure starts with the same modules as used for the FICTS normalization. The additional step is that a canonical tautomer form is determined (see below). This structure is defined as the FICuS parent structure, whose hash code becomes the FICuS identifier, thus yielding a tautomer-invariant representation of the input structure. Since, e.g., different salt forms, differently isotopically labeled variants, and different stereoisomers of a compound are usually seen by chemists as different chemicals, whereas different tautomers drawn for the same compound are not, FICuS is probably the closest chemoinformatics representation among our identifiers of how a chemist perceives a chemical.

The uuuuu parent structure and identifier are a much generalized representation of the input structure. During the normalization of the uuuuu parent structure only the largest organic fragment is kept, i.e. in the case of (organic) salts and coordination compounds any counterions or the metal complex center, respectively, are removed. The input structure is “un-charged” to its most reasonable state. Finally any information about stereochemistry and isotope labels is deleted from the structure. The uuuuu parent structure and identifier are therefore useful to link together closely related forms of the same chemical compound.

All three variants of parent structure and identifier are calculated when a structure is registered in CSDB and stored in the database. This gives one quite fine-grained control over how each chemical compound present in CSDB can be represented as well as searched with different degrees of sensitivity to different chemical features.

Figure 4 illustrates the relationships between the different parent structures and identifiers after structure normalization.

CSDB currently stores 70.6 million parent structures that are unique by their tautomer-invariant FICuS identifier. Each FICuS parent structure is linked to one or more tautomer-sensitive FICTS parent structures, each of which is in turn linked to one or more original structure records. The current count of FICTS parent structures and original structure records in CSDB is 72.0 million and 103.5 million, respectively. The generic uuuuu parent structure links together different FICuS parent structures that are highly related to each other in the way described above. A total of 65.3 million uuuuu parent structures are currently stored in CSDB.

Enumeration of tautomers

If a structure normalization procedure is to include handling of possible tautomerism of small molecules, several components need to be in place: (1) a set of rules for the possible molecular transforms that define the scope of what is meant by “tautomerism” in the context of this approach; (2) a practical implementation of the generation of tautomers, such as exhaustive enumeration of all unique tautomers within reasonable limits, e.g. achieved by setting certain program parameters; (3) definition of a canonical tautomer, e.g. based on a scoring scheme for various chemical features present in one tautomer vs. another. This latter point is essential if a tautomer-invariant connectivity-based identifier is to be calculated for each input structure, such as is done with the NCI/CADD Chemical Structure Identifiers.

Tautomer rules (SMIRKS transforms)

For the enumeration of tautomers, CACTVS uses a set of 21 tautomer rules that cover a wide range of typical 1,2-, 1,3-, 1,5-, 1,7-, 1,9- and 1,11 hydrogen atom shifts. The transforms encoded in these tautomer rules are based on the SMIRKS line notation originally developed by Daylight Chemical Information Systems, Inc., for the description of reaction substructures and the transformation of atoms and bonds during reactions [22]. Table 1 lists the 21 rules and their SMIRKS transforms used by CACTVS for the tautomer generation.

The SMIRKS in rule 1 and 2 address 1,3- and 1,5-keto-enol tautomerism of ketones and enols. Both rules are not restricted to keto and hydroxy groups but also include their sulfur, selenium end tellurium counterparts.

Rule 3 in Table 1 describes 1,3 hydrogen atom shifts of aliphatic imines. Rule 4 handles the special case of imines where a pyridine-type aromatic ring system is created or undone by an aliphatic hydrogen acceptor or donor carbon atom adjacent to the ring (atom 1 of rule 4).

The next seven rules cover hydrogen atom shifts on aromatic heterosystems or aliphatic heteroatoms. The first rule of this group, rule 5, addresses a special case of a short-range 1,3 hydrogen atom shift operation. The rule creates or undoes a heteroaromatic system if the central carbon atom (atom 2 in rule 5) is member of a ring system with six pi electrons. This constraint avoids the generation of unlikely high-energy tautomer forms in other ring systems. For rule 5, atom 1 must be a nitrogen atom, atom 2 has to be a carbon atom, and atom 3 has to be a nitrogen or oxygen atom.

Rule 6 handles 1,3 hydrogen migrations on aromatic hetero systems and aliphatic heteroatoms, however with fewer restrictions than the previous rule. In contrast to rule 5, here the central atom 2 is allowed to be a nitrogen or phosphorus atom, and for the two heteroatom positions 1 and 3 sulfur, oxygen, selenium, and tellurium atoms are additionally accepted.

The next five tautomer rules (rule 7–11) deal with long-range 1.5 hydrogen shifts (rule 7 and 8), or very long range hydrogen atom migrations across 7, 9, or 11 atoms (rule 9–11). While rule 8 is restricted to aromatic systems with nitrogen, oxygen, or sulfur atoms, rule 7 addresses specific aliphatic structures with selenium and tellurium as additional atom types. Rule 9–11 are quite similar to each other in what they do and vary only in the number of intermediate carbon atoms and the specification of element types at the terminal heteroatom positions.

The remaining rules 12–21 handle the tautomerism of very specific compound classes, functional groups and molecules. Rule 12 addresses the tautomerism occurring for furanones. This includes furanone-like molecules with a nitrogen or sulfur atom at terminal atom position 2. The interconversion between a keten and ynol group is governed by rule 13. This rule additionally accepts a sulfur, selenium or tellurium atom in place of the oxygen atom. The tautomerism of nitro groups defined in ionic form or with a pentavalent nitrogen atom is handled by rules 14 and 15. Rule 16 manages the tautomerism of simple oxim-nitroso groups; rule 17 handles the special case of oxim-nitroso tautomerism via a phenol system. The final group of rules (rule 18–21) addresses the tautomerism of cyanic/isocyanic acids, formamidinesulfinic acids, and phosphonic acids.

One type of tautomerism that was not included in this study is ring-chain tautomerism. To the best of our knowledge, no chemoinformatics tool in its standard implementation currently handles general ring-chain tautomerism, presumably because ring-chain tautomerism possesses—at least in the general case—more of a 3D nature than most other forms of tautomerism. I.e. while this type of tautomerism may be well-defined for, e.g., the standard carbohydrates, it is much less clear where, and whether at all, it can occur for any type of molecule for which, e.g., steric hindrance may prevent ring closing.

For the generation of all tautomers of a chemical compound, the SMIRKS rules in Table 1 have to be applied systematically to the structure, i.e. each side of each transform scheme has to be tested for a possible match to the structure and, if the match is successful, transformed to the other side. This has to be repeated systematically in case a new tautomeric center has been created by the previous step and the repeated application of the same transform or the application of another transform would generate yet another tautomer of the structure. If several SMIKRS transforms match the structure all possible combinations of tautomer transformations have to be executed during each step. This process has to be continued until no additional new tautomers can be found, a previously specified maximum number of tautomers has been generated, a specified maximum of transform operations has been performed, or a specified timeout is reached (though the latter was not used in this study).

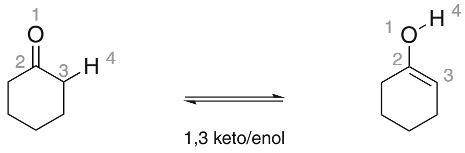
Implementation of the tautomer generation process in CACTVS

The CACTVS command *ens transform* generates all tautomers of a structure when applied to a so-called molecular ensemble (in which a structure is stored in CACTVS). The command returns the full set of possible tautomers for this

Table 1 SMIRKS transforms for the enumeration of tautomers. CACTVS provides an extended set of attributes for the definition of SMIRKS that have no counterpart in the original SMIRKS syntax, e.g. the attribute z*n* indicates the number *n* of heteroatoms substituted to the corresponding carbon atom. Another additional attribute in CACTVS is the e*n* attribute used in rule 5 (e6 on atom 2) which indicates that the corresponding carbon atom has to be member of a ring with at least *n* pi atoms

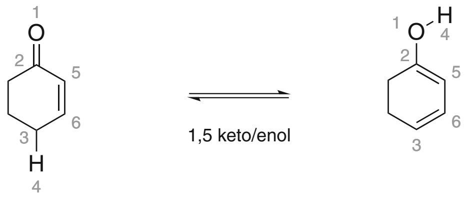
Rule 1: 1,3 (thio)keto/(thio)enol

[O,S,Se,Te;X1:1]=[C;z{1-2}:2][CX4R{0-2}:3]**[#1:4]** **>>[#1:4]**[O,S,Se,Te;X2:1][#6;z{1-2}:2]=[C,cz{0-1}R{0-1}:3]



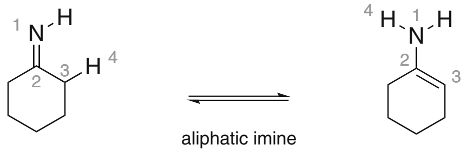
Rule 2: 1,5 (thio)keto/(thio)enol

[O,S,Se,Te;X1:1]=[Cz1H0:2][C:5]=[C:6][CX4z0,NX3:3]**[#1:4]**>> **[#1:4]**[O,S,Se,Te;X2:1][Cz1:2]=[C:5][C:6]=[Cz0,N:3]



Rule 3: simple (aliphatic) imine

[#1,a:5][NX2:1]=[Cz1:2][CX4R{0-2}:3]**[#1:4]**>>[#1,a:5][NX3:1](**[#1:4]**)[Cz1,Cz2:2]=[C:3]



Rule 4: special imine

[Cz0R0X3:1]([C:5])=[C:2][Nz0:3]**[#1:4]>>[#1:4]**[Cz0R0X4:1]([C:5])[c:2]=[nz0:3]

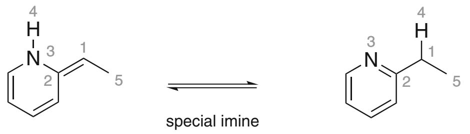
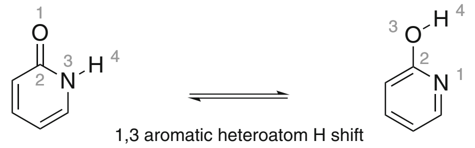


Table 1 continued

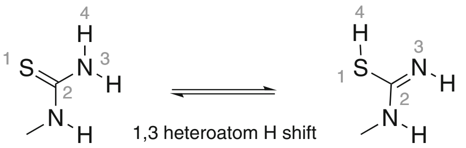
Rule 5: 1,3 aromatic heteroatom H shift

[#1:4][N:1][C;e6:2]=[O,NX2:3]>>[NX2,nX2:1]=[C,c;e6:2][O,N:3]**[#1:4]**



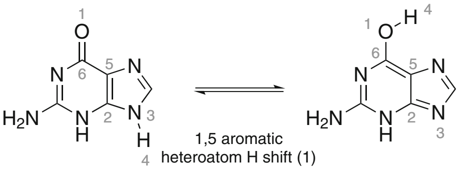
Rule 6: 1,3 heteroatom H shift

[N,n,S,s,O,o,Se,Te:1]=[NX2,nX2,C,c,P,p:2][N,n,S,O,Se,Te:3]**[#1:4]**>>**[#1:4]**[N,n,S,O,Se,Te:1][NX2,nX2,C,c,P,p:2]=[N,n,S,s,O,o,Se,Te:3]



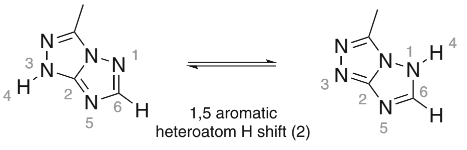
Rule 7: 1,5 (aromatic) heteroatom H shift (1)

[nX2,NX2,S,O,Se,Te:1]=[C,c,nX2,NX2:6][C,c:5]=[C,c,nX2:2][N,n,S,s,O,o,Se,Te:3]**[#1:4]**>>**[#1:4]**[N,n,S,O,Se,Te:1][C,c,nX2,NX2:6]=[C,c:5][C,c,nX2:2]=[NX2,S,O,Se,Te:3]



Rule 8: 1,5 aromatic heteroatom H shift (2)

[n,s,o:1]=[c,n:6][c:5]=[c,n:2][n,s,o:3]**[#1:4]>>[#1:4]**[n,s,o:1][c,n:6]=[c:5][c,n:2]=[n,s,o:3]

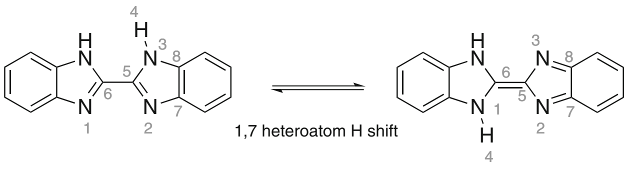


Rule 9: 1,7 (aromatic) heteroatom H shift

[nX2,NX2,S,O,Se,Te,Cz0X3:1]=[c,C,NX2,nX2:6][C,c:5]=[C,c,NX2,nX2:2][C,c,NX2,nX2:7]=[C,c,NX2,nX2:8][N,n,S,s,O,o,Se,Te:3]**[#1:4]**>>**[#1:4]**

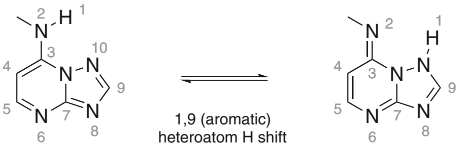
Table 1 continued

[N,n,S,O,Se,Te,Cz0X4:1][C,c,NX2,nX2:6]=[C,c:5][C,c,NX2,nX2:2]=[C,c,NX2,nX2:7][C,c,NX2,nX2:8]=[NX2,S,O,Se,Te:3][C,c,NX2,nX2:8]=[NX2,S,O,Se,Te:3]



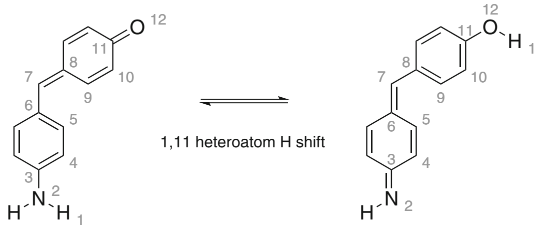
Rule 10: 1,9 (aromatic) heteroatom H shift

[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2:5]=[c,nX2:6][c,nX2:7]=[c,nX2:8][c,nX2,C:9]=[n,N,O:10]>>[N,n,O:2]=[C,c,nX2:3][c,nX2:4]=[c,nX2:5][c,nX2:6]=[c,nX2:7][c,nX2:8]=[c,nX2:9][n,O:10]**[#1:1]**



Rule 11: 1,11 (aromatic) heteroatom H shift

[#1:1][n,N,O:2][c,nX2,C:3]=[c,nX2,C:4][c,nX2:5]=[c,C,nX2:6][c,C,nX2:7]=[c,C,nX2:8][c,nX2,C:9]=[c,C,nX2:10][c,C,nX2:11]=[nX2,NX2,O:12]>>[NX2,nX2,O:2]=[C,c,nX2:3][c,C,nX2:4]=[c,C,nX2:5][c,C,nX2:6]=[c,C,nX2:7][c,C,nX2:8]=[c,C,nX2:9][c,C,nX2:10]=[c,C,nX2:11][nX2,O:12]**[#1:1]**



Rule 12: furanones

[#1:1][O,S,N:2][c,C;z2;r5:3]=[C,c;r5:4][c,C;r5:5]>>[O,S,N:2]=[Cz2r5:3][C&r5R{0-2}:4](**[#1:1]**)[C,c;r5:5]

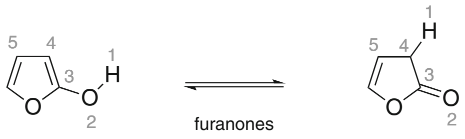
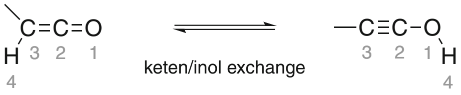


Table 1 continued

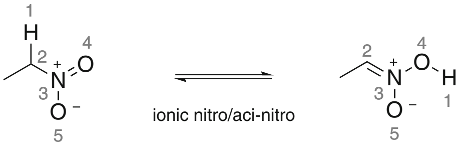
Rule 13: keten/ynol exchange

[O,S,Se,Te;X1:1]=[C:2]=[C:3]**[#1:4]**>>**[#1:4]**[O,S,Se,Te;X2:1][C:2]#[C:3]



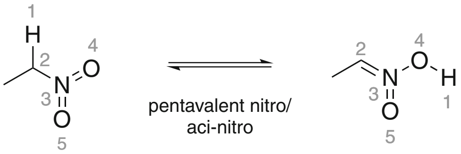
Rule 14: ionic nitro/aci-nitro

[#1:1][C:2][N+:3]([O–:5])=[O:4]>>[C:2]=[N+:3]([O-:5])[O:4]**[#1:1]** checkcharges



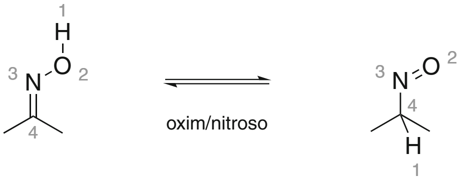
Rule 15: pentavalent nitro/aci-nitro

[#1:1][C:2][N:3](=[O:5])=[O:4]>>[C:2]=[N:3](=[O:5])[O:4]**[#1:1]**



Rule 16: oxim/nitroso

[#1:1][O:2][Nz1:3]=[C:4]>>[O:2]=[Nz1:3][C:4]**[#1:1]**



Rule 17: oxim/nitroso via phenol

[#1:1][O:2][N:3]=[C:4][C:5]=[C:6][C:7]=[O:8]>>[O:2]=[N:3][c:4]=[c:5][c:6]=[c:7][O:8]**[#1:1]**

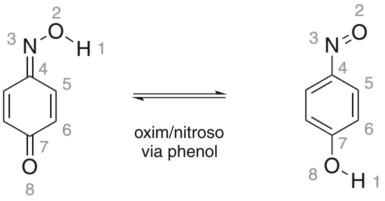
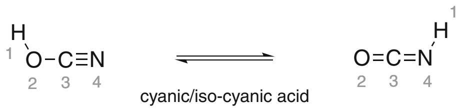


Table 1 continued

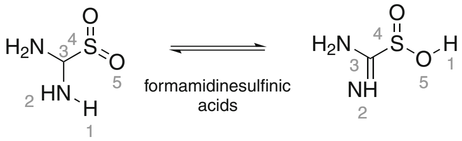
Rule 18: cyanic/iso-cyanic acids

[#1:1][O:2][C:3]#[N:4]>>[O:2]=[C:3]=[N:4]**[#1:1]**



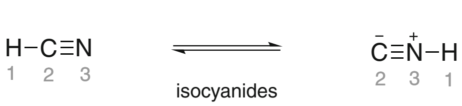
Rule 19: formamidinesulfinic acids

[#1:1][O,N:2][C:3]=[S,Se,Te:4]=[O:5]>>
[O,N:2]=[C:3][S,Se,Te:4][O:5]**[#1:1]**



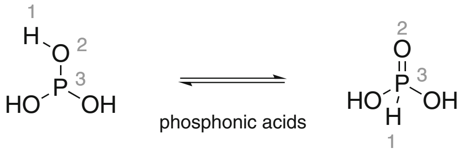
Rule 20: isocyanides

[#1:1][C0:2]#[N0:3]>>[C–:2]#[N+:3]**[#1:1]**
checkcharges checkaro



Rule 21: phosphonic acids

[#1:1][O:2][P:3]>>[O:2]=[P:3]**[#1:1]**



structure as a list of CACTVS molecular ensemble objects, each of which holding a single tautomer. The generation of all possible tautomers is accomplished by a systematic application of all SMIRKS transforms listed in Table 1. The way the transforms are applied is controlled by several parameters of the *ens transform* command, which can actually be used to perform any formal reaction that can be described as SMIRKS.

The specific parameters used for the tautomer generation during the calculation of our NCI/CADD Structure Identifiers are *direction*, *reactionmode*, *selectionmode*, and *maxstructures* or *maxtransforms,* which will be described in some detail because they can have significant influence on the generated results especially for larger and tautomerically more complicated molecules.

The first parameter, *direction*, is set to the value *bidirectional,* which means CACTVS attempts to match and execute each of the SMIRKS transforms in Table 1 in both possible directions of the formal reaction they describe. The parameter *reactionmode*, determining how multiple occurrences of the transform substructures in the original structure are handled, is used with the value *multistep*. This value specifies that a systematic application of all transforms is performed. Therefore, each new tautomer that has been generated by the application of one of the SMIRKS transforms is resubmitted again until no further new tautomers can be found.

The parameter *selectionmode* is set to the value *all*. This mode specifies that all SMIRKS transform in Table 1 are applied to any of the structures generated by any previous step of the tautomer generation, not just to the molecular ensembles obtained by the previous step and in the strict order the transforms have been specified (as would be the case with the mode value *sequence*).

The parameter *maxstructures* specifies the maximum number of tautomers that should be returned by the *ens transform* command. For some structures, the enumeration of tautomers runs into a combinatorial explosion of generated tautomer structures. For the calculation of our NCI/CADD Structure Identifiers, we set *maxstructures* to an upper limit of 1,000.

Because of the exhaustive application of the SMIRKS rules, in most cases at least a subset of tautomers resulting from a specific rule is identical to already generated tautomers. For the de-duplication of generated tautomer structures, the algorithm behind the *ens transform* command filters any tautomer duplicates by calculating one of the hash code variants available in CACTVS (E_HASHISY [21]) for each tautomer, thus confining the final set of generated tautomers to a unique set of structures.

If the limit of 1,000 generated tautomers has been reached before exhaustive application of the transform rules, the tautomer generation process is terminated and the corresponding identifier is then flagged as (possibly) unreliable. Such cases of a very high number of generated tautomers are mostly the result of long, complex sequences of transforms that result in tautomer structures of only minor practical interest. Analyses we performed, however, showed that for the majority of structures registered in the database the canonical form was reached within these limits. As mentioned above, for the calculation of our NCI/CADD Structure Identifiers it is not the entire set of tautomers that is of actual interest but instead to obtain one canonical tautomer. While determined by definition (since true, energy-based stability calculations can not be performed, as discussed above), such a canonical tautomer should obviously strive to be a very plausible structure by all accepted measures. Therefore, there is a high likelihood for the canonical tautomer to be found among the first 1,000 generated structures; i.e. even if a structure identifier is nominally flagged as unreliable after tautomer generation there is a high probability that it represents the correct canonical tautomer.

In addition to the program parameters described so far, several other *ens transform* command flags that have an influence on the enumerated tautomers are used. The flags *checkaro* is set globally (i.e. for all transforms), which undoes the special CACTVS modification of the original SMIRKS definition, i.e. to consider uppercase elements (in the parlance of the SMILES/SMIRKS syntax) as undefined with respect to aromaticity in a substructure definition, and reverts to the original Daylight implementation insofar as uppercase elements can only match aliphatic atoms, while lowercase elements can only match aromatic atoms. The second flag, *preservecharges*, controls whether a matched atom is changed to the charge of the matching atom in the specified SMIRKS transform. By default, CACTVS performs this change of charges as long as the corresponding atom has sufficient electrons. If the *preservecharges* flag is set, charges are not modified. This flag affects rules 14 and 20. For rule 14, the *preservecharges* flag is un-set since charges should be modified by this transform. For both rules 14 and 20 an additional flag, *checkcharges*, is set, which specifies that the number of formal charges on the matching side of the transform must be identical to the number of charges on the matched structure.

Definition of the canonical tautomer

After all tautomers of a given input structure have been enumerated, a canonical tautomer has to be defined among this set of generated tautomers. As mentioned, defining the truly chemically preferred tautomer is difficult since it requires treatment of effects such as dipole–dipole repulsion, electronic, and thermodynamic effects and is even quite likely dependent, for the same set of tautomers, on factors such as solvent, temperature, basic *vs*. acidic environment, etc. The influence of all these effects cannot be calculated easily and quickly. Therefore, we implemented a fast, empirical, rule-based rating algorithm in CACTVS instead. This rating system was established by analyzing several different sets of tautomers and the known preferred tautomer members included in these data sets. Table 2 shows the scoring rules obtained from this analysis.

Each scoring rule is based on the occurrence of certain structure fragment. The general scoring of a tautomer is increased or decreased by the number of scoring points of the corresponding fragment multiplied by its number of occurrences.

The tautomer that has garnered the best scoring of all tautomers in the set is defined as the canonical tautomer of the given structure. If more than one tautomer gets the maximum scoring, the tautomer with the largest hash code value is, quite arbitrarily from a structural point of view, selected as the canonical tautomer form. Generally, this approach does not guarantee that the tautomer defined as the canonical one is the chemically most reasonable or the lowest in energy in absolute terms; however, this is not

Table 2 Scoring of structure fragments used for the definition of a canonical tautomer

Structure fragment	Scoring points
Each carbocyclic aromatic ring	+150
Each aromatic ring	+100
Each benzoquinones (including imine and thio analogs, [C]1([C]=[C][C]([C]=[C]1)=,:[N,S,O])=,:[N,S,O], penalize cyclohexanetetrone-like structures)	+25
Each oxim group (C=N[OH])	+4
Each double bond between a carbon atom (C) and an oxygen atom (O)	+2
Each double bond between a nitrogen atom (N) and an oxygen atom (O)	+2
Each double bond between a phosphorus atom (P) and an oxygen atom (O)	+2
Each non-aromatic double bond between a carbon atom (C) and a heteroatom (X)	+1
Each methyl group (penalize structures with terminal double bonds)	+1
Each guanidine group with a double bond on the terminal nitrogen atom (NC(=N)[N][!H])	+1
Each guanidine group with an endocyclic double bond ([N;R][C;R]([N])=[N;R])	+2
Each P-H, S-H, Se-H and Te-H bond	−1
Each aci-nitro group (C=N(=O)[OH])	−4

The scoring points were obtained by an analysis of different sets of tautomers including the known preferred tautomer needed here. The more important aspect is to always find the same tautomer form as the endpoint of the described enumeration process regardless of which tautomer form was given as the starting point to the algorithms. This is guaranteed if the list of SMIRKS transforms was applied exhaustively during the enumeration of tautomers. Even if the limit of 1,000 generated tautomers was hit, the algorithm displayed a still very high reliability of generating the true canonical tautomer (had exhaustive enumeration been done) for compounds of the sizes typically found in small molecule databases.

Normalization of stereochemistry in the canonical tautomer

As shown above (Fig. 2), stereochemistry of a chemical compound can be affected by tautomerism. In order to tackle this problem, we expanded the existing algorithm in CACTVS for defining the canonical tautomer by adding a step for the correction of stereochemistry. Figure 5 illustrates how this works for the example of methyl propenyl ketone.

Methyl propenyl ketone can be drawn as an E or Z stereoisomer, and both stereoisomers can be separated spectroscopically and have different CAS Registry numbers [23, 24]. Notwithstanding this, CACTVS creates a common set of formal tautomers in which the location of the original double bond—including its stereochemistry—has changed.

As mentioned above, whether structures such as these two stereoisomers of methyl propenyl ketone actually interconvert depends on conditions including pH, temperature, and solvent, and in general on structural effects such as steric hindrance [25], conformer energy differences, and barriers to internal rotation [26]. However, all these effects are way beyond the scope of chemical structure identifiers or a database registration process that should be usable for millions of compounds. Another aspect is that, arguably for aesthetic reasons, it is quite common for chemist to draw double bonds with unspecified or unknown stereochemistry in the E form. Therefore, for our tautomer-invariant FICuS parent structure and identifier, we decided to disregard stereochemistry on double bonds that do not have a fixed location during tautomer generation. In contrast to this, the tautomer-sensitive FICTS parent structure and identifier preserve both the specific tautomer and any stereochemistry on double bonds even if it could change position because of tautomerism.

We also developed an extension of the algorithm that removes stereochemistry assignment in a similar way for sp3 hybridized atoms that have assigned R/S stereochemistry but changed to an sp2 hybridized atom at least once during tautomer generation (see the thalidomide example in Fig. 2). For atoms of this type, racemization may occur because of tautomerism. However, general application of this module would be problematic. For instance, the R and S forms of amino acids would not be distinguishable by our tautomer-invariant FICuS identifier anymore since one formal tautomeric form contains an sp2 hybridized alpha carbon atom. We therefore currently do not use this module in our structure normalization algorithm.

An illustration of our exhaustive tautomer enumeration is shown in Fig. 6 for the example of 2-hydroxy-3,4-dimethoxy-6-methylbenzaldehyde (**1**).

CACTVS generates 12 additional tautomers (**2**-**13**). They are displayed in Fig. 6, which also shows the transform rule (from the set given in Table 1) that led to each interconversion listed, plus each tautomer's scoring calculated on the basis of the scoring scheme elaborated in Table 2. The tautomer that received the highest scoring among the 13 tautomers was structure **1**. It is therefore regarded as the canonical tautomer, which is used as the tautomer representing the FICuS parent structure. One

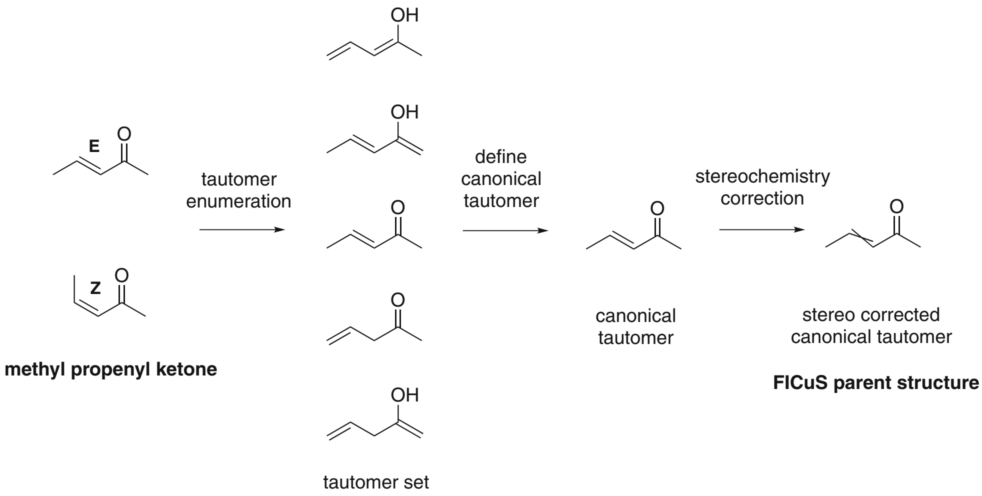


Fig. 5 Normalization of stereochemistry for the canonical tautomer, involving double bonds whose stereochemistry is disregarded in the final step producing the canonical tautomer when the original stereo bond does not have a fixed location during tautomer generation

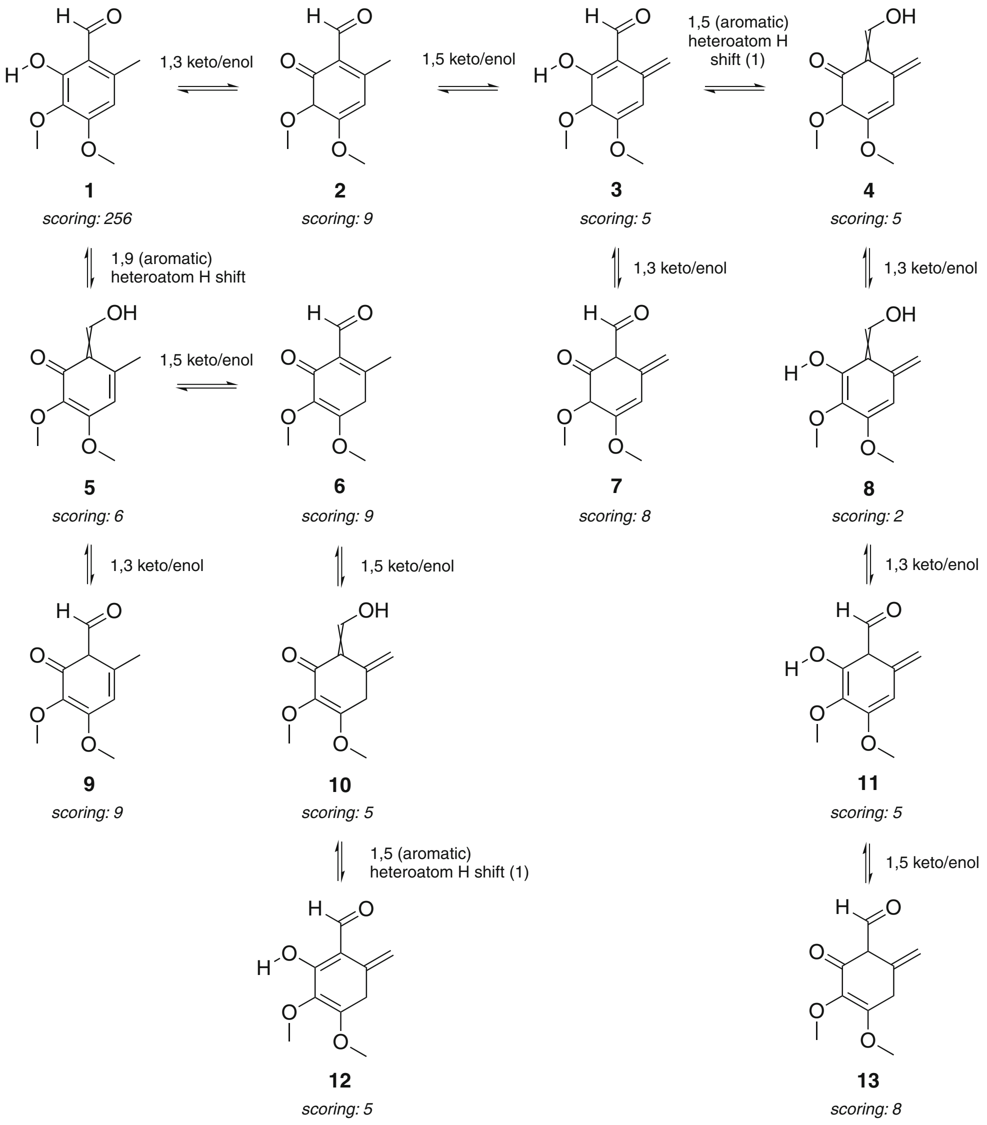


Fig. 6 Enumeration of all tautomers of 2-hydroxy-3,4-dimethoxy-6-methylbenzaldehyde (**1**)

should note that a number of double bonds are drawn as crossed bonds. This indicates that these bonds have been explicitly assigned "no stereochemistry" because, though generated in the specific tautomer, they are mobile thus do not have a fixed stereochemistry throughout the entire enumeration process.

Results

Tautomeric analysis of CSDB

Tautomeric overlap within each database ("local" overlap)

Table 3 lists all releases of original databases that are currently contained in CSDB, with various counts and percentages including the results of the analyses based on tautomeric overlap found within each individual database in CSDB.

As can be seen, the majority of individual databases were downloaded from PubChem. As source we used the full database dump provided as Substance SD files at PubChem's FTP server [17]. This download was performed on 26-Jun-2007, and a second time on 10-Jun-2008 to update CSDB with substance records that had not been part of the first download. This also added a handful of entirely new databases that were only present in this second download.

For ChemNavigator's *iResearch Library*, we followed the CADD Group's quarterly update of this database. The latest update that was included for this paper is the July 2009 release of the *iRL*. Structure records coming from *the iRL* are registered on the basis of ChemNavigator's Structure ID, which is their unique structure identifier. It should be noted that even for Structure IDs that

Table 3 List of all databases present in CSDB. With: the original publisher and the source from where the database was obtained; the number of original structure records and the unique structure counts after de-duplication by the FICTS and FICuS identifiers, and the percentage of duplicate structures by both identifiers; the difference between unique structure counts between the FICTS and FICuS parent structure sets

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Structure record count	Unique structure count (FICTS)	Unique structure count (FICuS)	% Duplicates by FICTS	% Duplicates by FICuS	Discrepancy FICTS/FICuS structure count	% Duplicates FICTS-FICuS
ACD 3D	MDL/Symyx	MDL/Symyx	1999-01-01	221,661	215,154	214,614	2.93	3.17	540	0.24
ACX	CambridgeSoft	CambridgeSoft	1999-12-31	137,001	101,009	100,729	26.27	26.47	280	0.20
Ambinter	Ambinter	PubChem	2008-06-10	2,692,132	2,688,795	2678,948	0.12	0.48	9,847	0.36
Aronis	Aronis	PubChem	2008-06-10	23,389	23,389	23,385	0.0	0.01	4	0.01
Asinex	Asinex	PubChem	2007-07-26	362,469	362,464	362,464	<0.01	0.0	0	0.0
Asinex Building Blocks	Asinex	Asinex	2005-04-01	5,248	5,248	5,248	0.0	0.0	0	0.0
Asinex Gold Collection	Asinex	Asinex	2006-06-01	227,479	227,475	227,475	<0.01	0.0	0	0.0
Asinex Platinum Collection	Asinex	Asinex	2006-06-01	130,646	130,646	130,646	0.0	0.0	0	0.0
BIND	BIND	PubChem	2007-07-26	1,207	1,205	1,203	0.16	0.33	2	0.17
BindingDB	BindingDB	PubChem	2007-07-26	8,492	8,470	8,458	0.25	0.4	12	0.15
			2008-06-10	12,747	12,705	12,699	0.32	0.37	6	0.05
BioByte QSAR	BioByte	BioByte	2006-05-01	155,296	154,679	153,801	0.39	0.96	878	0.57
BioCyc	BioCyc	PubChem	2007-07-26	1,660	1,307	1,285	21.26	22.59	22	1.33
Biosynth	Biosynth	PubChem	2008-06-10	2,079	1,934	1,931	6.97	7.11	3	0.14
Calbiochem	Calbiochem	PubChem	2008-06-10	1,665	1,591	1,591	4.44	4.44	0	0.0
CambridgeSoft	CambridgeSoft	PubChem	2007-07-26	10,458	10,143	10,120	3.01	3.23	23	0.22
CC PMLSC	CC PMLSC	PubChem	2007-07-26	222	217	217	2.25	2.25	0	0.0
			2008-06-10	173	173	172	0.0	0.57	1	0.57
ChEBI	ChEBI	PubChem	2007-07-26	8,767	8,373	8,238	4.49	6.03	135	1.54
			2008-06-10	2,604	2,563	2,515	1.57	3.41	48	1.84
ChemBank	ChemBank	PubChem	2007-07-26	413,586	339,066	338,520	18.01	18.15	546	0.14
			2008-06-10	1,194,169	1014,374	1011,147	15.05	15.32	3,227	0.27
ChemBlock	ChemBlock	PubChem	2007-07-26	107,570	107,290	107,192	0.26	0.35	98	0.09
ChemBridge	ChemBridge	PubChem	2007-07-26	433,971	433,970	433,970	<0.01	0.0	0	0.0
ChemBridge 100 k Lib	ChemBridge	ChemBridge	2002-02-01	100,000	99,997	99,920	<0.01	0.08	77	0.07
ChemDB	ChemDB	PubChem	2007-07-26	3,564,882	3,549,580	3501,958	0.42	1.76	47,622	1.34
			2008-06-10	2	2	2	0.0	0.0	0	0.0
ChemDiv Diversity Collection	ChemDiv	ChemDiv	2004-09-01	495466	495,455	495,395	<0.01	0.01	60	0.01
ChemExper Chemical Directory	ChemExper Chemical Directory	PubChem	2007-07-26	156,258	156,113	155,698	0.09	0.35	415	0.26
ChemSpider	ChemSpider	PubChem	2008-06-10	17,064,543	16,871,574	16,537,474	1.13	3.08	334,100	1.95

Table 3 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Structure record count	Unique structure count (FICTS)	Unique structure count (FICuS)	% Duplicates by FICTS	% Duplicates by FICuS	Discrepancy FICTS/FICuS structure count	% Duplicates FICTS-FICuS
CHMIS-C	UMich	UMich	2004-11-01	8,572	8,022	7,992	6.41	6.76	30	0.35
CMC	MDL/Symyx	MDL/Symyx	2006-01-01	8,757	8,742	8,732	0.17	0.28	10	0.11
CMLD-BU	CMLD-BU	PubChem	2007-07-26	1,629	1,619	1,619	0.61	0.61	0	0.0
Columbia University Molecular Screening Center	Columbia University Molecular Screening Center	PubChem	2008-06-10	399	391	391	2.0	2.0	0	0.0
ComGenex	ComGenex	ComGenex	2006-03-01	184,266	184,266	184,266	0.0	0.0	0	0.0
ComGenex Unique Reagents	ComGenex	ComGenex	2006-03-01	330	330	330	0.0	0.0	0	0.0
Diabetic Complications Screening	Diabetic Complications Screening	PubChem	2007-07-26	1,040	1	1	99.9	99.9	0	0.0
DiscoveryGate	Symyx	PubChem	2007-07-26	4,608,993	4,602,080	4,581,587	0.14	0.59	20,493	0.45
			2008-06-10	1,261,853	1,260,435	1,260,101	0.11	0.13	334	0.02
DrugBank	DrugBank	PubChem	2008-06-10	4,763	4,419	4,409	7.22	7.43	10	0.21
Dupont Library	Dupont	MDDP/NCI	2004-04-01	179,008	174,977	174,745	2.25	2.38	232	0.13
Emory University Molecular Libraries Screening Center	Emory University Molecular Libraries Screening Center	PubChem	2007-07-26	101,567	101,540	101,523	0.02	0.04	17	0.02
			2008-06-10	4,367	4,335	4,333	0.73	0.77	2	0.04
EPA DSSTox	EPA DSSTox	PubChem	2007-07-26	4,258	4,103	4,101	3.64	3.68	2	0.04
			2008-06-10	12,950	6,635	6,630	48.76	48.8	5	0.04
Exchemistry	Exchemistry	PubChem	2008-06-10	2,057	2,057	2,057	0.0	0.0	0	0.0
FDA CDER Chronic/ Subchronic	FDA/CDER	FDA.CDER	2006-05-01	84	84	84	0.0	0.0	0	0.0
FDA CDER Genetox	FDA/CDER	FDA/CDER	2006-05-01	231	181	181	21.64	21.64	0	0.0
FDA CFSAN Genetox	FDA/CFSAN	FDA/CFSAN	2006-05-01	487	400	400	17.86	17.86	0	0.0
FDA Genet/Reprod/ Carcino	FDA/CDER	FDA/CDER	2006-01-01	6,912	6,820	6,810	1.33	1.47	10	0.14
InFarmatik	InFarmatik	PubChem	2008-06-10	1,077	1,077	1,077	0.0	0.0	0	0.0
iResearch Library	ChemNavigator	ChemNavigator	2004-07-01	13,352,734	13,350,546	13,323,974	0.01	0.21	26,572	0.20
			2004-10-01	263,325	263,311	261,858	<0.01	0.55	1,453	0.55
			2005-01-01	5,036,400	5,036,247	5,035,543	<0.01	0.01	704	0.01
			2005-04-01	480,405	480,402	480,350	<0.01	0.01	52	0.01
			2005-07-01	479,324	479,305	479,289	<0.01	<0.01	16	<0.01

Table 3 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Structure record count	Unique structure count (FICTS)	Unique structure count (FICuS)	% Duplicates by FICTS	% Duplicates by FICuS	Discrepancy FICTS/FICuS structure count	% Duplicates FICTS-FICuS
			2005-10-01	377,246	377,238	376,265	<0.01	0.26	973	0.25
			2006-01-01	210,500	210,497	210,490	<0.01	<0.01	7	<0.01
			2006-04-01	580,332	580,328	579,878	<0.01	0.07	450	0.07
			2006-07-01	4,218,127	4,218,019	4,217,935	<0.01	<0.01	84	<0.01
			2006-10-01	220,602	220,561	220,322	0.01	0.12	239	0.11
			2007-01-01	4,045,672	4,045,656	4,045,492	<0.01	<0.01	164	<0.01
			2007-07-01	333,578	333,568	333,561	<0.01	<0.01	7	<0.01
			2007-10-01	3,148,102	3,148,066	3,148,008	<0.01	<0.01	58	<0.01
			2008-01-01	528,308	528,294	528,280	<0.01	<0.01	14	<0.01
			2008-04-01	536,482	536,477	536,302	<0.01	0.03	175	0.03
			2008-07-01	548,386	548,373	548,335	<0.01	<0.01	38	<0.01
			2008-10-01	380,396	380,394	380,299	<0.01	0.02	95	0.02
			2009-01-01	564,140	564,133	564,082	<0.01	0.01	51	<0.01
			2009-04-01	784,758	784,731	784,343	<0.01	0.05	388	0.04
			2009-07-01	22,225,547	22,219,140	22,211,624	0.02	0.06	7,516	0.04
Jubilant Kinase Inhibitors	Jubilant	Jubilant	2004-12-01	170,000	163,799	163,518	3.64	3.81	281	0.17
KEGG	KEGG	PubChem	2007-07-26	16,938	14,313	14,233	15.49	15.97	80	0.48
			2008-06-10	3,551	2,477	2,475	30.24	30.30	2	0.06
KUMGM	KUMGM	PubChem	2007-07-26	3,349	3,109	3,107	7.16	7.22	2	0.06
Leadscope FDA	Leadscope/FDA	PubChem	2007-07-26	724	588	588	18.78	18.78	0	0.0
LifeChem Building Blocks	LifeChem	LifeChem	2006-05-01	4,027	4,027	4,020	0.0	0.17	7	0.17
LifeChem Stock Compounds	LifeChem	LifeChem	2006-05-01	204,955	204,954	204,765	<0.01	0.09	189	0.09
LifeChem Virtual Compounds	LifeChem	LifeChem	2006-05-01	179,649	179,648	179,648	<0.01	0.0	0	0.0
LipidMAPS	LipidMAPS	PubChem	2007-07-26	10,128	9,628	9,590	4.93	5.31	38	0.38
			2008-06-10	308	284	284	7.79	7.79	0	0.0
MDDR	MDL/Symyx	MDL/Symyx	2006-03-01	165,595	164,666	164,561	0.56	0.62	105	0.06
MDL Patent Database	MDL/Symyx	MDL/Symyx	2005-11-01	38,363	30,980	30,840	19.24	19.61	140	0.37
MDL Toxicity Database	MDL/Symyx	MDL/Symyx	2005-11-01	147,308	147,144	147,006	0.11	0.2	138	0.09
MDPI	MDPI	MDPI	2004-11-01	10,655	10,513	10,478	1.33	1.66	35	0.33
MICAD	MICAD	PubChem	2007-07-26	188	187	187	0.53	0.53	0	0.0
			2008-06-10	76	76	76	0.0	0.0	0	0.0

Table 3 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Structure record count	Unique structure count (FICTS)	Unique structure count (FICuS)	% Duplicates by FICTS	% Duplicates by FICuS	Discrepancy FICTS/FICuS structure count	% Duplicates FICTS-FICuS
MLSMR	MLSMR	PubChem	2007-07-26	207,811	204,326	204,198	1.67	1.73	128	0.06
			2008-06-10	75,904	75,460	75,453	0.58	0.59	7	<0.01
MMDB	MMDB	PubChem	2007-07-26	65,388	9,450	9,373	85.54	85.66	77	0.12
			2008-06-10	26,470	4,456	4,428	83.16	83.27	28	0.11
MOLI	MOLI	PubChem	2007-07-26	1,951	1,774	1,774	9.07	9.07	0	0.0
MTDP/NCI	MTDP/NCI	PubChem	2007-07-26	106,186	106,016	105,931	0.16	0.24	85	0.08
			2008-06-10	301	301	301	0.0	0.0	0	0.0
NatChemBio	NatChemBio	PubChem	2007-07-26	1,565	1,447	1,446	7.53	7.6	1	0.07
			2008-06-10	871	844	841	3.09	3.44	3	0.35
NCGC	NCGC	PubChem	2007-07-26	54,728	53,714	53,703	1.85	1.87	11	0.02
			2008-06-10	13,797	10,992	10,971	20.33	20.48	21	0.15
NCI open database	NCI/DTP	NCI/CADD	2006-07-01	263,465	253,957	253,550	3.6	3.76	407	0.16
		PubChem	2007-07-26	268,696	251,396	250,854	6.43	6.64	542	0.21
			2008-06-10	5,383	5,372	5,365	0.2	0.33	7	0.13
NCI-NP	NCI/DTP	NCI/DTP	2002-02-01	124,700	120,736	119,587	3.17	4.10	1,149	0.93
NIAID HIV/OI	NIAID	NIAID	2006-02-01	138,693	133,064	132,806	4.05	4.24	258	0.19
		PubChem	2007-07-26	155,624	149,609	149,341	3.86	4.03	268	0.17
			2008-06-10	4,260	4,068	4,052	4.5	4.88	16	0.38
NIH Clinical Collection	NIH Clinical Collection	PubChem	2008-06-10	489	473	472	3.27	3.47	1	0.20
NINDS-ADSP	NINDS	PubChem	2007-07-26	1,040	1,033	1,031	0.67	0.86	2	0.19
NINDS-PANACHE	NINDS	PubChem	2007-07-26	10	10	10	0.0	0.0	0	0.0
NIST MS-Lib	NIST	NIST	2006-01-01	177,495	171,241	170,917	3.52	3.7	324	0.18
		PubChem	2007-07-26	177,495	171,239	170,920	3.52	3.7	319	0.18
NIST WebBook	NIST	NIST	2006-01-01	54,146	51,621	51,451	4.66	4.97	170	0.31
		PubChem	2007-07-26	54,125	51,573	51,403	4.71	5.02	170	0.31
NLM ChemIDplus	NLM	NLM	2006-03-01	269,276	255,638	255,138	5.06	5.25	500	0.19
		PubChem	2007-07-26	383,789	274,146	273,597	28.56	28.71	549	0.15
NMMLSC	NMMLSC	PubChem	2007-07-26	5,776	5,770	5,770	0.1	0.1	0	0.0
			2008-06-10	438	438	438	0.0	0.0	0	0.0
NMRShiftDB	NMRShiftDB	PubChem	2007-07-26	19,414	19,016	18,956	2.05	2.35	60	0.30
NTP-CHSD	NTP	NTP	1991-08-01	1,588	1,383	1,383	12.9	12.9	0	0.0
NTP-PTC	NTP	NTP	2002-09-01	417	401	401	3.83	3.83	0	0.0

Table 3 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Structure record count	Unique structure count (FICTS)	Unique structure count (FICuS)	% Duplicates by FICTS	% Duplicates by FICuS	Discrepancy FICTS/FICuS structure count	% Duplicates FICTS-FICuS
ORST Small Molecule Screening Center	ORST Small Molecule Screening Center	PubChem	2008-06-10	1,998	1,996	1,993	0.1	0.25	3	0.15
PASS Training Set	LSFBDD/IBMC/RAMS	LSFBDD/IBMC/ RAMS	2006-02-01	60,620	60,487	60,172	0.21	0.73	315	0.52
PCMD	PCMD	PubChem	2007-07-26	27	27	27	0.0	0.0	0	0.0
			2008-06-10	65	65	65	0.0	0.0	0	0.0
PDSP	PDSP	PubChem	2007-07-26	2,871	2,868	2,867	0.1	0.13	1	0.03
ProbeDB	ProbeDB	PubChem	2007-07-26	10	1	1	90.0	90.0	0	0.0
			2008-06-10	273	1	1	99.63	99.63	0	0.0
Prous Science Drugs of the Future	Prous Science Drugs of the Future	PubChem	2007-07-26	4,426	4,421	4,417	0.11	0.2	4	0.09
			2008-06-10	202	202	202	0.0	0.0	0	0.0
R&D Chemicals	R&D Chemicals	PubChem	2008-06-10	8,352	8,352	8,352	0.0	0.0	0	0.0
RTECS	NIOSH/CDC	NIOSH/CDC	2004-06-01	144,729	137,216	137,094	5.19	5.27	122	0.08
SDCCG	SDCCG	PubChem	2007-07-26	54,838	54,597	54,565	0.43	0.49	32	0.06
SDCCG	SDCCG	PubChem	2008-06-10	1,215	1,175	1,172	3.29	3.53	3	0.24
SGC-Ox	SGC-Ox	PubChem	2007-07-26	319	311	308	2.5	3.44	3	0.94
SGC-Sto	SGC-Sto	PubChem	2007-07-26	17	17	17	0.0	0.0	0	0.0
Shanghai Institute of Organic Chemistry	Shanghai Institute of Organic Chemistry	PubChem	2008-06-10	3,080	2,443	2,428	20.68	21.16	15	0.48
Sigma–Aldrich	Sigma–Aldrich	PubChem	2007-07-26	56,080	37,534	37,519	33.07	33.09	15	0.02
SMID	SMID	PubChem	2007-07-26	7,161	6,516	6,500	9.0	9.23	16	0.23
Southern Research Institute—HTS	Southern Research Institute—HTS	PubChem	2008-06-10	1,114	1,114	1,113	0.0	0.08	1	0.08
Specs	Specs	PubChem	2007-07-26	205,958	205,954	205,954	0.0	0.0	0	0.0
SRMLSC	SRMLSC	PubChem	2008-06-10	304	304	304	0.0	0.0	0	0.0
Structural Genomics Consortium	Structural Genomics Consortium	PubChem	2007-07-26	87	87	87	0.0	0.0	0	0.0
			2008-06-10	90	90	90	0.0	0.0	0	0.0
The Scripps Research Institute Molecular Screening Center	The Scripps Research Institute Molecular Screening Center	PubChem	2007-07-26	2	2	2	0.0	0.0	0	0.0
			2008-06-10	16,231	16,185	16,180	0.28	0.31	5	0.03
Thomson Pharma	Thomson Pharma	PubChem	2007-07-26	2,303,463	2,285,548	2,277,301	0.77	1.13	8,247	0.36

Table 3 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Structure record count	Unique structure count (FICTS)	Unique structure count (FICuS)	% Duplicates by FICTS	% Duplicates by FICuS	Discrepancy FICTS/FICuS structure count	% Duplicates FICTS-FICuS
			2008-06-10	224,196	223,873	223,630	0.14	0.25	243	0.11
Total TOSLab Building Blocks	Total TOSLab Building Blocks	PubChem	2007-07-26	910	910	909	0.0	0.1	1	0.1
UM-BBD	UM-BBD	PubChem	2007-07-26	1,081	1,067	1,062	1.29	1.75	5	0.46
			2008-06-10	38	38	38	0.0	0.0	0	0.0
University of Pittsburgh Molecular Library Screening Center	University of Pittsburgh Molecular Library Screening Center	PubChem	2008-06-10	303	273	273	9.9	9.9	0	0.0
UPCMLD	UPCMLD	PubChem	2007-07-26	2,111	1,879	1,879	10.99	10.99	0	0.0
			2008-06-10	495	493	493	0.4	0.4	0	0.0
USAMRIID In Silico-Screened Structures	USAMRIID	USAMRIID	2004-06-01	376,062	359,673	359,554	4.35	4.38	119	0.03
WDI	Derwent/Thomson Reuters	Derwent/Thomson Reuters	2006-02-01	79,618	69,439	69,283	12.78	12.98	156	0.20
Web of Science	Web of Science	PubChem	2007-07-26	20	20	18	0.0	10.0	2	10.0
Wombat 2005.02	Sunset Molecular Discovery	Sunset molecular discovery	2005-02-01	135,673	120,475	120,287	11.2	11.34	188	0.14
xPharm	xPharm	PubChem	2007-07-26	2,462	2,137	2,135	13.2	13.28	2	0.08
ZINC	ZINC	PubChem	2007-07-26	3,813,885	3,748,592	3,707,913	1.71	2.77	40679	1.06

ChemNavigator has marked as inactive or not available any more, we keep them registered in CSDB. Our main interest of having these structures (that were declared as being available at least at some point in time) available for *in silico* screening experiments overrides the consideration whether this structures are currently available or not (which needs to be individually ascertained before actual sample orders anyway).

ChemNavigator's structure records currently represent approximately 56% of all structure records registered in CSDB, the percentage of PubChem records is approximately 38%, the remaining original database combine to approximately 6%. Some of these databases, especially some US Government databases such as the Open NCI database, the NIST WebBook, or the NLM ChemIDplus set had been in our collection for a long time (obtained directly as SD files from the original sources) and were not jettisoned for this study, therefore their substantial overlap with the same database's release from PubChem is not surprising. The large difference in record counts between "our" NLM ChemIDplus version and the ChemIDplus set coming from PubChem stems from the fact that this database contains a lot of records that have no structure. PubChem registered all these records with their own Substance ID, whereas we added only those records that contained a structure in the original ChemIDplus SD file.

Table 3 lists the unique structure counts obtained after FICTS structure normalization (tautomer-sensitive) and FICuS structure normalization (tautomer-invariant), respectively, for each database release in CSDB, as well as the percentage of duplicates with regard to the number of original structure records calculated from these counts. As Table 3 shows, the major part of de-duplication is already achieved by the FICTS structure normalization, which does not include any tautomer normalization. The average percentage of duplicates found across all databases during this normalization step is approximately 6.8% (average of all values in column "*% Duplicates by FICTS"* in Table 3) when compared to the number of original structure records. For the tautomer-invariant FICuS identifier, the average percentage of duplicates found is at about 7.0% (average of all values in column "*% Duplicates by FICuS"*), i.e. the average difference between the de-duplication steps by FICTS structure normalization and FICuS normalization is surprisingly small for each release. Especially ChemNavigator seems to use a very strong algorithm for the normalization of structures in general, which also appears to include a very strong tautomer de-duplication step. For the numbers for all database releases obtained from PubChem, it is important to remember that we used the "raw" substance files which had not undergone any normalization by PubChem, thus these numbers represent the quality of the original database releases.

If the tautomer-invariant FICuS identifier hash code value for a chemical compound has a different value than the tautomer-sensitive FICTS identifier, this means that neither the original structure record nor the FICTS parent structure are identical to the canonical tautomer as represented by the FICuS parent structure. In the entire CSDB database, this occurred for 8.9% (9,224,751 records) of the 103,497,350 original structure records. Based on the number of unique FICTS parent structures (72,034,119 records), 8.6% of the FICTS parent structures (6,198,011 records) changed to a different tautomer during the FICuS normalization procedure.

From the perspective of unique FICuS parent structures, about 98.5% of them (69,561,639 records) had a one-to-one relationship to a FICTS parent structure, i.e. even though the tautomer-invariant FICuS parent structure may represent a different tautomer than the tautomer-sensitive FICTS parent structure, there were no conflicts in the sense that any other original tautomer structures (FICTS parent structures) were found assigned to this same canonical tautomer (FICuS parent structure). The group of FICuS parent structures with this one-to-one relationship to a FICTS parent structure represents about 96.6% of the FICTS parent structures and 95.2% of the 103,497,350 original structure records, respectively.

The remaining 1,078,853 FICuS parent structures (1.5%) had multiple FICTS parent structures assigned to them, which is an indication of a tautomer conflict. The frequency of such conflicts is not simply a function of the individual database size: The numbers in the last column of Table 3 range from exactly 0 to nearly 2%.This finding argues against the possible objection that our tautomerism definition is too "aggressive" and will therefore hit a certain percentage of structures in any database no matter how carefully that database was processed or curated.

These tautomer conflicts can be grouped into three classes: (a) the number of tautomer-sensitive FICTS parent structures assigned to one tautomer-invariant FICuS parent structure exceeded the number of original databases in which the FICuS parent structure occurred, (b) the number of FICTS parent structures was the same as the number of databases in which the corresponding FICuS parent structure occurred, or (c) the number of FICTS parent structure was smaller than the number of database. The explanations for these cases are: (a) tautomer conflicts occurred for a specific chemical compound across all databases, plus some tautomer conflicts occurred even within a single database, (b) the same chemical compound was represented as different tautomers in all databases but there were no conflicts within a particular database, and (c), there were several database groups such that each database in its group consistently shared one tautomer representation for a specific chemical compound with all other group members, but

Table 4 Number of tautomer conflicts. As conflict is regarded a case when a tautomer-invariant parent structure (FICuS) is assigned to more than one tautomer-sensitive parent structures (FICTS)

	FICuS parent structure		FICTS parent structure		Original structure record	
	Count	%	Count	%	Count	%
(a)	119,632	0.17	285,502	0.40	315,277	0.30
(b)	398,079	0.56	877,182	1.22	1,013,198	0.98
(c)	561,142	0.79	1,334,800	1.85	3,590,508	3.47
sum	1,078,853	1.52	2,497,484	3.47	4,918,983	4.75

There are three types of such cases: (a) tautomer conflicts occur for a specific chemical compound in one or more databases with conflicts even among structure records of a single database, (b) the same chemical compound is represented as different tautomers in different database but there are no conflicts within each single database, and (c), there were several groups of databases which each consistently shared one tautomer representation for a specific chemical compound, but the different database groups used different tautomer representations

different database groups used different tautomer representations. Table 4 shows the different parent structure counts for these cases, listing the number of unique structures by FICuS structure normalization (column "*FICuS parent structure count* ") and the number of FICTS parent structures that are linked to structures regarded as unique by FICuS structure normalization (column "*FICTS parent structure count*"). Analogous numbers are shown for the count of structure records in the original databases (column "*original structure record count*" in Table 4). The percentage values are calculated with respect to the corresponding unique structure or record counts in Fig. 4.

In Fig. 7 we show the tautomeric situation we found for the compound 1-phenyl-3-methyl-4-benzoyl-pyrazolone-5 (HPMBP) as a "real-life" example of a case in class (a). HPMBP is used in liquid membranes for the selective removal of metal ions or molecules from dilute solution [27]. The selectivity and efficiency for the extraction of metal ions seems to depend specifically on the tautomeric form of HPMBP, which in turn is dependent on solvent and the concentration of HPMBP in the liquid membrane. The structures **14-20** shown in Fig. 7 represent all formal tautomers enumerated by CACTVS. The interconversion between the tautomers in subset **14-17** seems to be energetically unhindered [27].

For five of the seven tautomers, a pair of stereoisomers can be drawn; the remaining two are achiral structures. Three of the tautomers have a CAS Registry Number (CASRN) assigned. CASRN 4551-69-3 has 859 references assigned in SciFinder, CASRN 33064-14-1 has 49 references and CASRN 12711-31-1 occurs with 3 references, which seems to be an indication that the first two structures are the most important tautomeric forms. Our algorithm in CACTVS defines **15** as the canonical tautomer.

Figure 7 also shows the occurrence counts of the HPMBP tautomers in CSDB. One can see that the majority of the seven formally possible tautomeric representations were actually found in one or more of the constituent databases of CSDB. Tautomer **14** was found in six databases (Ambinter, ChemDB, ChemSpider, DiscoveryGate, iResearch Library, Thomson Pharma), tautomer **15** in 12 databases (ACD 3D, Ambinter, BindingDB, ChemBank, ChemDB, ChemSpider, iResearch Library, MLSMR, NIAID HIV/OI, Scripps Research Institute Molecular Screening Center, Thomson Pharma, ZINC), tautomer **16** was found without indication of stereo configuration in 16 databases (ACD 3D, ACX,

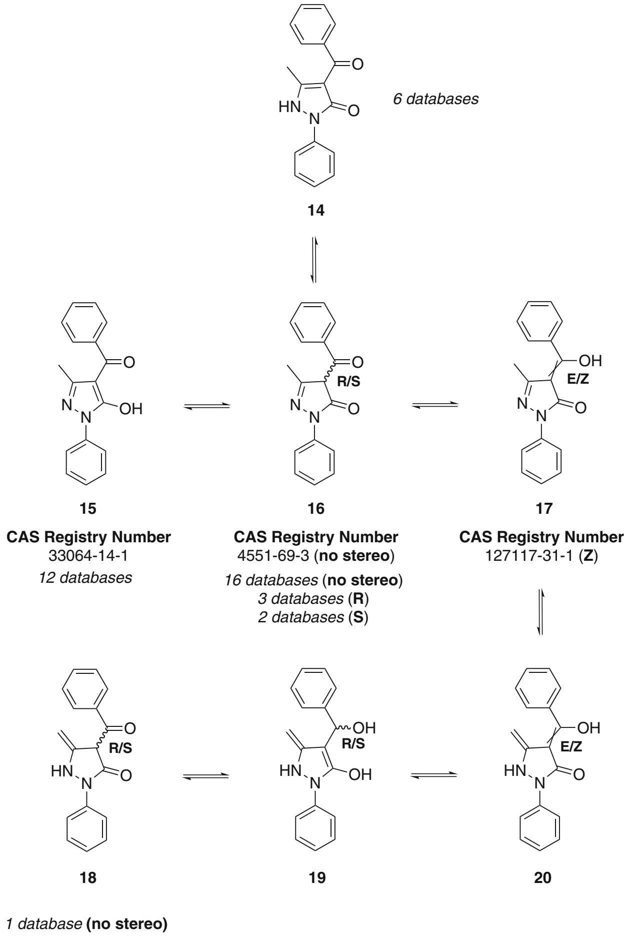


Fig. 7 Example of a tautomer conflict found for 1-phenyl-3-methyl-4-benzoyl-pyrazolone-5 (HPMBP)

Ambinter, BioByte QSAR, ChemBank, ChemBridge, ChemDB, ChemSpider, DiscoveryGate, EPA GCES, MLSMR, NCI Open Database, NIST MS-Lib, NLM ChemIDplus, Sigma–Aldrich and Thomson Pharma), as R stereoisomer in three databases (ChemSpider, ECOTOX, and ZINC) and as S stereoisomer in two databases (ChemSpider and ZINC). Finally, tautomer **18** was found in ChemDB with no stereo information present.

Tautomeric overlap across databases in CSDB (“global” overlap)

Table 5 provides the results of a more “global” analysis of tautomerism in CSDB in the sense that we look at tautomeric multiplicity of each structure across all of the databases in CSDB and not just within each database (release) as it was done for Table 3.

The column “*FICuS structures with formal tautomerism*” lists the numbers and percentages of canonical tautomer structures (FICuS structure set) for which CACTVS generates at least one additional formal tautomer. The average percentage value of structures showing tautomerism in CSDB according to our (admittedly “aggressive”) definition is 68.3%. The next column, “*occurrences of FICuS structures with multiple FICTS assignment*” gives the numbers of FICuS parent structures that occurred with a global conflict, i.e. had more than one FICTS parent structure assigned somewhere in CSDB (see Table 4). The average percentage of FICuS structures for which this occurs in each database release of the CSDB is 9.5%. The last column in Table 5 lists the numbers and percentages of FICuS parent structures which occurred exclusively in that one database release. By definition, this is the fraction of structures for which it is not possible to have a tautomer conflict with other databases in CSDB.

Systematic enumeration of tautomers

Another aspect we were interested in was how large the “chemical space of formal tautomers” is that can be enumerated from the structures in CSDB according to complete set of tautomeric transform rules available in Table 1. By applying these rules to the set of 70,640,491 canonical tautomers (FICuS parent structure set), all possible tautomers were enumerated.

We used essentially the same setup as for the calculation of our NCI/CADD Structure Identifiers, for which we set a maximum of 1,000 tautomers to be generated per input structure. However, since CACTVS attempts to systematically generate tautomers until 1,000 structurally *different* tautomers are found, it can occur that CACTVS has to perform a large fraction of all possible combinations of SMIRKS transform until this limit is reached (or all possible combinations have been exhausted). For molecules with many tautomeric centers, this can take in the range of minutes. To limit CPU time to a manageable level for this experiment, we therefore set a limit of 1,000 transforms for each structure. This limit is typically reached earlier than the 1,000 tautomer limit, and thus leads to a more linear scaling of CPU time as a function of the number of tautomeric centers.

This procedure created a set of 680,556,829 chemical structures including the original FICuS parent structure set. Table 6 shows how often each CACTVS transform rule from Table 1 was used in the creation of this tautomer set. This may provide a useful statistics about the prevalence—and thus importance in algorithmic approaches—of the various tautomeric transforms encountered for a real database, not just assessed on theoretical grounds. As one can see, the distribution varies widely, and ranges from an order of 100 to more than 100 million.

The number of FICuS parent structures for which the generation of tautomers was not exhaustive (because the limit of 1,000 transforms had been reached) was approximately 1,2 million (~1.7%).

Table 7 shows the distribution of tautomers generated for all canonical tautomers (FICuS parent structures). For only 13.8% of the FICuS parent structures did the application of our rules not generate any tautomers. The maximum number of tautomers generated for one structure was 832. The majority of structures (62.7%) had between one and ten tautomers.

It bears repeating at this point that our definition of tautomerism is a quite “aggressive” one, i.e. quite a few of the tautomers generated by the SMIRKS rules described earlier would be regarded by a chemist as a minor or even entirely unlikely form. For instance, even if a molecule contains as the sole functional group only one single amide group, its imidic acid form is generated as a possible tautomer. Therefore, the number of molecules not showing any appreciable tautomerism—could the experiment be conducted for 70+ million substances—is in all likelihood quite a bit underestimated in Table 7. It must be emphasized, however, that our approach is not meant to most faithfully represent the experimental situation. Its main purpose is not to avoid any energetically unfavorable forms but to tie together as many of the conceivable tautomeric representations of a compound as possible. Our experience has shown that one has to expect to encounter any formally possible tautomer when working with many different databases and tens of millions of structures, and handling this correctly requires a systematic and comprehensive enumeration of tautomers. For this purpose, it is of no negative consequence if we equate tautomeric forms with each other among which there are high-energy forms that are not likely to exist under normal conditions.

Table 5 Analysis of "global" tautomerism in CSDB

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Unique structure count (FICuS)	FICuS parent structures with formal tautomerism		Occurrences of FICuS parent structures with multiple FICTS parent structure assignment		FICuS parent structures exclusive to the database release	
					Count	%	Count	%	Count	%
ACD 3D	MDL/Symyx	MDL/Symyx	1999-01-01	214,614	124,829	58.16	30,373	14.15	21,623	10.08
ACX	CambridgeSoft	CambridgeSoft	1999-12-31	100,729	54,666	54.27	6,810	6.76	8,217	8.16
Ambinter	Ambinter	PubChem	2008-06-10	2,678,948	2,064,072	77.05	25,186	0.94	277,080	10.34
Aronis	Aronis	PubChem	2008-06-10	23,385	20,347	87.01	0	0.0	2,731	11.68
Asinex	Asinex	PubChem	2007-07-26	362,464	291,172	80.33	7	<0.01	73,463	20.27
Asinex Building Blocks	Asinex	Asinex	2005-04-01	5,248	3,262	62.16	0	0.0	608	11.59
Asinex Gold Collection	Asinex	Asinex	2006-06-01	227,475	174,579	76.75	7	<0.01	39,093	17.19
Asinex Platinum Collection	Asinex	Asinex	2006-06-01	130,646	115,075	88.08	10	0.01	34,174	26.16
BIND	BIND	PubChem	2007-07-26	1,203	921	76.56	31	2.58	226	18.79
BindingDB	BindingDB	PubChem	2007-07-26	8,458	7,068	83.57	30	0.35	1,163	13.75
			2008-06-10	12,699	10,528	82.9	3,956	31.15	825	6.5
BioByte QSAR	BioByte	BioByte	2006-05-01	153,801	91,447	59.46	52,936	34.42	10,235	6.65
BioCyc	BioCyc	PubChem	2007-07-26	1,285	931	72.45	108	8.4	275	21.4
Biosynth	Biosynth	PubChem	2008-06-10	1,931	1,160	60.07	356	18.44	218	11.29
Calbiochem	Calbiochem	PubChem	2008-06-10	1,591	1,184	74.42	275	17.28	225	14.14
CambridgeSoft	CambridgeSoft	PubChem	2007-07-26	10,120	6,213	61.39	1,408	13.91	619	6.12
CC PMLSC	CC PMLSC	PubChem	2007-07-26	217	168	77.42	0	0.0	24	11.06
			2008-06-10	172	157	91.28	0	0.0	54	31.4
ChEBI	ChEBI	PubChem	2007-07-26	8,238	4,266	51.78	684	8.3	921	11.18
			2008-06-10	2,515	1,307	51.97	310	12.33	245	9.74
ChemBank	ChemBank	PubChem	2007-07-26	338,520	254,500	75.18	3,632	1.07	37,414	11.05
			2008-06-10	1,011,147	816,927	80.79	405,912	40.14	69,909	6.91
ChemBlock	ChemBlock	PubChem	2007-07-26	107,192	78,845	73.55	0	0.0	17,959	16.75
ChemBridge	ChemBridge	PubChem	2007-07-26	433,970	322,423	74.3	201	0.05	53,926	12.43
ChemBridge 100 k Lib	ChemBridge	ChemBridge	2002-02-01	99,920	73,884	73.94	2	<0.01	17,760	17.77
ChemDB	ChemDB	PubChem	2007-07-26	3,501,958	2,592,957	74.04	59,112	1.69	468,965	13.39
			2008-06-10	2	0	0.0	2	100.0	0	0.0
ChemDiv Diversity Collection	ChemDiv	ChemDiv	2004-09-01	495,395	385,876	77.89	71,699	14.47	7,345	1.48
ChemExper Chemical Directory	ChemExper Chemical Directory	PubChem	2007-07-26	155,698	93,621	60.13	20,170	12.95	458	0.29
ChemSpider	ChemSpider	PubChem	2008-06-10	16,537,474	11,418,636	69.05	930,014	5.62	5,577,994	33.73

Table 5 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Unique structure count (FICuS)	FICuS parent structures with formal tautomerism		Occurrences of FICuS parent structures with multiple FICTS parent structure assignment		FICuS parent structures exclusive to the database release	
					Count	%	Count	%	Count	%
CHMIS-C	UMich	UMich	2004-11-01	7,992	4,455	55.74	493	6.17	3,526	44.12
CMC	MDL/Symyx	MDL/Symyx	2006-01-01	8,732	5,818	66.63	950	10.88	235	2.69
CMLD-BU	CMLD-BU	PubChem	2007-07-26	1,619	1,240	76.59	19	1.17	13	0.8
Columbia University Molecular Screening Center	Columbia University Molecular Screening Center	PubChem	2008-06-10	391	198	50.64	14	3.58	153	39.13
ComGenex	ComGenex	ComGenex	2006-03-01	184,266	152,569	82.8	6,579	3.57	1,482	0.8
ComGenex unique reagents	ComGenex	ComGenex	2006-03-01	330	257	77.88	70	21.21	1	0.3
Diabetic Complications Screening	Diabetic Complications Screening	PubChem	2007-07-26	1	0	0.0	0	0.0	0	0.0
DiscoveryGate	Symyx	PubChem	2007-07-26	4,581,587	3,029,660	66.13	246,048	5.37	10,503	0.23
			2008-06-10	1,260,101	993,999	78.88	41,697	3.31	57,180	4.54
DrugBank	DrugBank	PubChem	2008-06-10	4,409	3,062	69.45	605	13.72	450	10.21
Dupont Library	Dupont	MDDP/NCI	2004-04-01	174,745	107,268	61.39	10,644	6.09	28,654	16.4
Emory University Molecular Libraries Screening Center	Emory University Molecular Llibraries Screening Center	PubChem	2007-07-26	101,523	79,264	78.07	13,471	13.27	42	0.04
			2008-06-10	4,333	3,029	69.91	700	16.16	152	3.51
EPA DSSTox	EPA DSSTox	PubChem	2007-07-26	4,101	1,998	48.72	600	14.63	1	0.02
			2008-06-10	6,630	2,870	43.29	770	11.61	23	0.35
Exchemistry	Exchemistry	PubChem	2008-06-10	2,057	1,465	71.22	96	4.67	0	0.0
FDA CDER Chronic/Subchronic	FDA/CDER	FDA/CDER	2006-05-01	84	56	66.67	15	17.86	0	0.0
FDA CDER Genetox	FDA/CDER	FDA/CDER	2006-05-01	181	125	69.09	30	16.57	0	0.0
FDA CFSAN Genetox	FDA/CFSAN	FDA/CFSAN	2006-05-01	400	216	54.0	66	16.5	2	0.5
FDA Genet/Reprod/Carcino	FDA/CDER	FDA/CDER	2006-01-01	6,810	3,368	49.46	788	11.57	1,165	17.11
InFarmatik	InFarmatik	PubChem	2008-06-10	1,077	519	48.19	195	18.11	1	0.09
iResearch Library	ChemNavigator	ChemNavigator	2004-07-01	13,323,974	10,687,768	80.21	512,426	3.85	9,623,698	72.23
			2004-10-01	261,858	204,119	77.95	45,095	17.22	104,306	39.83
			2005-01-01	5,035,543	4,255,548	84.51	105,694	2.1	4,328,401	85.96
			2005-04-01	480,350	412,308	85.83	12,008	2.5	324,134	67.48
			2005-07-01	479,289	402,593	84.0	26,610	5.55	259,516	54.15
			2005-10-01	376,265	307,360	81.69	14,113	3.75	190,820	50.71

Table 5 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Unique structure count (FICuS)	FICuS parent structures with formal tautomerism		Occurrences of FICuS parent structures with multiple FICTS parent structure assignment		FICuS parent structures exclusive to the database release	
					Count	%	Count	%	Count	%
			2006-01-01	210,490	173,675	82.51	16,300	7.74	34,843	16.55
			2006-04-01	579,878	476,119	82.11	21,960	3.79	429,856	74.13
			2006-07-01	4,217,935	3,534,703	83.8	60,074	1.42	3,456,568	81.95
			2006-10-01	220,322	161,823	73.45	7,892	3.58	115,867	52.59
			2007-01-01	4,045,492	3,565,754	88.14	19,237	0.48	3,878,268	95.87
			2007-07-01	333,561	271,602	81.42	5,797	1.74	245,861	73.71
			2007-10-01	3,148,008	2,805,813	89.13	17,595	0.56	2,865,405	91.02
			2008-01-01	528,280	415,137	78.58	7,759	1.47	412,591	78.1
			2008-04-01	536,302	366,459	68.33	5,087	0.95	438,680	81.8
			2008-07-01	548,335	404,730	73.81	3,900	0.71	519,290	94.7
			2008-10-01	380,299	295,198	77.62	1,690	0.44	366,496	96.37
			2009-01-01	564,082	373,572	66.23	8,674	1.54	436,089	77.31
			2009-04-01	784,343	440,348	56.14	6,879	0.88	735,868	93.82
			2009-07-01	22,211,624	21,995,486	99.03	91,672	0.41	21,859,016	98.41
Jubilant Kinase Inhibitors	Jubilant	Jubilant	2004-12-01	163,518	153,169	93.67	7,916	4.84	103,571	63.34
KEGG	KEGG	PubChem	2007-07-26	14,233	9,519	66.88	1,765	12.4	884	6.21
			2008-06-10	2,475	1,662	67.15	216	8.73	438	17.7
KUMGM	KUMGM	PubChem	2007-07-26	3,107	1,883	60.61	130	4.18	796	25.62
Leadscope FDA	Leadscope/FDA	PubChem	2007-07-26	588	345	58.67	98	16.67	0	0.0
LifeChem Building Blocks	LifeChem	LifeChem	2006-05-01	4,020	2,614	65.02	713	17.74	2	0.05
LifeChem Stock Compounds	LifeChem	LifeChem	2006-05-01	204,765	158,064	77.19	26,429	12.91	581	0.28
LifeChem Virtual Compounds	LifeChem	LifeChem	2006-05-01	179,648	138,146	76.9	6,405	3.57	176	0.1
LipidMAPS	LipidMAPS	PubChem	2007-07-26	9,590	7,926	82.65	207	2.16	1,062	11.07
			2008-06-10	284	174	61.27	1	0.35	180	63.38
MDDR	MDL/Symyx	MDL/Symyx	2006-03-01	164,561	124,584	75.71	7,020	4.27	69,148	42.02
MDL Patent Database	MDL/Symyx	MDL/Symyx	2005-11-01	30,840	20,239	65.63	1,959	6.35	12,647	41.01
MDL Toxicity Database	MDL/Symyx	MDL/Symyx	2005-11-01	147,006	81,460	55.41	8,847	6.02	31,877	21.68
MDPI	MDPI	MDPI	2004-11-01	10,478	6,026	57.51	2,006	19.14	14	0.13
MICAD	MICAD	PubChem	2007-07-26	187	106	56.68	5	2.67	24	12.83
			2008-06-10	76	52	68.42	1	1.32	52	68.42

Table 5 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Unique structure count (FICuS)	FICuS parent structures with formal tautomerism		Occurrences of FICuS parent structures with multiple FICTS parent structure assignment		FICuS parent structures exclusive to the database release	
					Count	%	Count	%	Count	%
MLSMR	MLSMR	PubChem	2007-07-26	204,198	158,268	77.51	36,550	17.9	1,182	0.58
			2008-06-10	75,453	59,699	79.12	11,106	14.72	2,996	3.97
MMDB	MMDB	PubChem	2007-07-26	9,373	6,805	72.6	987	10.53	3,356	35.8
			2008-06-10	4,428	3116	70.37	488	11.02	1,651	37.29
MOLI	MOLI	PubChem	2007-07-26	1,774	1030	58.06	27	1.52	1,085	61.16
MTDP/NCI	MTDP/NCI	PubChem	2007-07-26	105,931	77,945	73.58	18,545	17.51	242	0.23
			2008-06-10	301	203	67.44	0	0.0	1	0.33
NatChemBio	NatChemBio	PubChem	2007-07-26	1,446	1,084	74.97	160	11.07	58	4.01
			2008-06-10	841	595	70.75	103	12.25	246	29.25
NCGC	NCGC	PubChem	2007-07-26	53,703	41,392	77.08	4,311	8.03	184	0.34
			2008-06-10	10,971	7,534	68.67	994	9.06	2,083	18.99
NCI Open Database	NCI/DTP	NCI/CADD	2006-07-01	253,550	150,496	59.36	28,804	11.36	29,119	11.48
		PubChem	2007-07-26	250,854	148,818	59.32	30,026	11.97	14,321	5.71
			2008-06-10	5,365	3,813	71.07	249	4.64	3,354	62.52
NCI-NP	NCI/DTP	NCI/DTP	2002-02-01	119,587	76,664	64.11	3,644	3.05	72,939	60.99
NIAID HIV/OI	NIAID	NIAID	2006-02-01	132,806	74,347	55.98	16,002	12.05	10,077	7.59
		PubChem	2007-07-26	149,341	88,058	58.96	18,212	12.19	10,495	7.03
			2008-06-10	4,052	3,064	75.62	591	14.59	110	2.71
NIH Clinical Collection	NIH Clinical Collection	PubChem	2008-06-10	472	333	70.55	85	18.01	3	0.64
NINDS-ADSP	NINDS	PubChem	2007-07-26	1,031	716	69.45	137	13.29	10	0.97
NINDS-PANACHE	NINDS	PubChem	2007-07-26	10	10	100.0	3	30.0	0	0.0
NIST MS-Lib	NIST	NIST	2006-01-01	170,917	75,311	44.06	13,615	7.97	3,295	1.93
		PubChem	2007-07-26	170,920	75,306	44.06	13,618	7.97	3,042	1.78
NIST WebBook	NIST	NIST	2006-01-01	51,451	16,897	32.84	2,482	4.82	547	1.06
		PubChem	2007-07-26	51,403	16,888	32.85	2,481	4.83	540	1.05
NLM ChemIDplus	NLM	NLM	2006-03-01	255,138	137,090	53.73	19,911	7.8	7,747	3.04
		PubChem	2007-07-26	273,597	147,403	53.88	21,223	7.76	8,915	3.26
NMMLSC	NMMLSC	PubChem	2007-07-26	5,770	4,637	80.36	916	15.88	0	0.0
			2008-06-10	438	354	80.82	17	3.88	25	5.71
NMRShiftDB	NMRShiftDB	PubChem	2007-07-26	18,956	7,729	40.77	1,584	8.36	1,810	9.55

Table 5 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Unique structure count (FICuS)	FICuS parent structures with formal tautomerism		Occurrences of FICuS parent structures with multiple FICTS parent structure assignment		FICuS parent structures exclusive to the database release	
					Count	%	Count	%	Count	%
NTP-CHSD	NTP	NTP	1991-08-01	1,383	504	36.44	202	14.61	87	6.29
NTP-PTC	NTP	NTP	2002-09-01	401	157	39.15	60	14.96	21	5.24
ORST Small Molecule Screening Center	ORST Small Molecule Screening Center	PubChem	2008-06-10	1,993	1,396	70.05	268	13.45	1	0.05
PASS Training Set	LSFBDD/IBMC/RAMS	LSFBDD/IBMC/ RAMS	2006-02-01	60,172	44,157	73.38	5,402	8.98	6,585	10.94
PCMD	PCMD	PubChem	2007-07-26	27	24	88.89	2	7.41	0	0.0
			2008-06-10	65	51	78.46	1	1.54	2	3.08
PDSP	PDSP	PubChem	2007-07-26	2,867	1,895	66.1	483	16.85	31	1.08
ProbeDB	ProbeDB	PubChem	2007-07-26	1	0	0.0	0	0.0	0	0.0
			2008-06-10	1	0	0.0	0	0.0	0	0.0
Prous Science Drugs of the Future	Prous Science Drugs of the Future	PubChem	2007-07-26	4,417	3,182	72.04	534	12.09	109	2.47
			2008-06-10	202	146	72.28	25	12.38	9	4.46
R&D chemicals	R&D chemicals	PubChem	2008-06-10	8,352	3,260	39.03	585	7.0	51	0.61
RTECS	NIOSH/CDC	NIOSH/CDC	2004-06-01	137,094	75,400	55.0	9,065	6.61	20,510	14.96
SDCCG	SDCCG	PubChem	2007-07-26	54,565	40,767	74.71	9,247	16.95	11	0.02
SDCCG	SDCCG	PubChem	2008-06-10	1,172	652	55.63	193	16.47	98	8.36
SGC-Ox	SGC-Ox	PubChem	2007-07-26	308	275	89.29	92	29.87	10	3.25
SGC-Sto	SGC-Sto	PubChem	2007-07-26	17	17	100.0	8	47.06	0	0.0
Shanghai Institute of Organic Chemistry	Shanghai Institute of Organic Chemistry	PubChem	2008-06-10	2,428	1,975	81.34	388	15.98	394	16.23
Sigma–Aldrich	Sigma–Aldrich	PubChem	2007-07-26	37,519	15,908	42.4	3,122	8.32	539	1.44
SMID	SMID	PubChem	2007-07-26	6,500	4,566	70.25	848	13.05	1,081	16.63
Southern Research Institute—HTS	Southern Research Institute—HTS	PubChem	2008-06-10	1,113	716	64.33	161	14.47	18	1.62
Specs	Specs	PubChem	2007-07-26	205,956	150,127	72.89	28,320	13.75	33	0.02
SRMLSC	SRMLSC	PubChem	2008-06-10	304	188	61.84	29	9.54	14	4.61
Structural Genomics Consortium	Structural Genomics Consortium	PubChem	2007-07-26	87	74	85.06	27	31.01	0.	0.0
			2008-06-10	90	77	85.56	25	27.78	0	0.0

Table 5 continued

Database name	Original publisher	Source/ downloaded from	Downloaded/ released at	Unique structure count (FICuS)	FICuS parent structures with formal tautomerism		Occurrences of FICuS parent structures with multiple FICTS parent structure assignment		FICuS parent structures exclusive to the database release	
					Count	%	Count	%	Count	%
The Scripps Research Institute Molecular Screening Center	The Scripps Research Institute Molecular Screening Center	PubChem	2007-07-26	2	2	100.0	0	0.0	0	0.0
			2008-06-10	16,180	10,609	65.57	2,663	16.46	50	0.31
Thomson Pharma	Thomson Pharma	PubChem	2007-07-26	2,277,301	1,389,970	61.04	132,517	5.82	129,218	5.67
			2008-06-10	223,630	145,692	65.15	9,605	4.3	97,628	43.66
Total TOSLab Building Blocks	Total TOSLab Building Blocks	PubChem	2007-07-26	909	666	73.27	244	26.84	0	0.0
UM-BBD	UM-BBD	PubChem	2007-07-26	1,062	579	54.52	198	18.64	57	5.37
			2008-06-10	38	21	55.26	6	15.79	12	31.58
University of Pittsburgh Molecular Library Screening Center	University of Pittsburgh Molecular Library Screening Center	PubChem	2008-06-10	273	248	90.48	67	24.54	0	0.0
UPCMLD	UPCMLD	PubChem	2007-07-26	1,879	1,415	75.31	38	2.02	25	1.33
			2008-06-10	493	418	84.79	10	2.03	91	18.46
USAMRIID In Silico-Screened Structures	USAMRIID	USAMRIID	2004-06-01	359,554	326,765	90.88	146	0.04	359,429	99.97
WDI	Derwent/Thomson Reuters	Derwent/Thomson Reuters	2006-02-01	69,283	49,238	71.07	6,459	9.32	8,996	12.98
Web of Science	Web of Science	PubChem	2007-07-26	18	9	50.0	2	11.11	0	0.0
Wombat 2005.02	Sunset Molecular Discovery	Sunset molecular discovery	2005-02-01	120,287	90,072	74.88	9,272	7.71	27,974	23.26
xPharm	xPharm	PubChem	2007-07-26	2,135	1,535	71.9	374	17.52	19	0.89
ZINC	ZINC	PubChem	2007-07-26	3,707,913	2,944,497	79.41	300,108	8.09	153,822	4.15

Table 6 Frequency of application of CACTVS transforms in the systematic generation of all tautomers for the FICuS parent structure (canonical tautomer) set

Transform rule	Generated tautomers	
	Count	%
Rule 1: 1.3 (thio)keto/(thio)enol	173,002,712	25.4
Rule 2: 1.5 (thio)keto/(thio)enol	11,541,452	1.7
Rule 3: simple (aliphatic) imine	3,5917,415	5.3
Rule 4: special imine	4,306,155	0.6
Rule 5: 1.3 aromatic heteroatom H shift	25,678,446	3.8
Rule 6: 1.3 heteroatom H shift	250,453,882	36.8
Rule 7: 1.5 (aromatic) heteroatom H shift (1)	27,542,770	4.0
Rule 8: 1.5 aromatic heteroatom H shift (2)	26,819	<0.1
Rule 9: 1.7 (aromatic) heteroatom H shift	57,242,472	8.4
Rule 10: 1.9 (aromatic) heteroatom H shift	5,061,731	0.7
Rule 11: 1.11 (aromatic) heteroatom H shift	1,374,235	0.2
Rule 12: furanones	17,860,604	2.6
Rule 13: keten/ynol exchange	57,989	<0.1
Rule 14: ionic nitro/aci-nitro	428,266	0.1
Rule 15: pentavalent nitro/aci-nitro	129	<0.1
Rule 16: oxim/nitroso	505,695	0.1
Rule 17: oxim/nitroso via phenol	131,502	0.2
Rule 18: cyanic/iso-cyanic acids	181	<0.1
Rule 19: formamidinesulfinic acids	1,392	<0.1
Rule 20: isocyanides	229	<0.1
Rule 21: phosphonic acids	54,926	<0.1

Table 7 : Distribution of the number of tautomers generated per FICuS parent structure

Canonical tautomers (FICuS parent structures) with	Count	%
no tautomers	9,756,186	13.8
one tautomer	10,721,845	15.2
2–10 tautomers	33,532,284	47.5
11–25 tautomers	10,870,312	15.4
25–50 tautomers	2,622,587	3.7
51–100 tautomers	1,136,066	1.6
101–200 tautomers	565,199	0.8
201–300 tautomers	104,875	0.1
301–400 tautomers	35,144	<0.1
401–500 tautomers	17,241	<0.1
501–600 tautomers	4,323	<0.1
601–700 tautomers	1,400	<0.1
701–800 tautomers	362	<0.1
801–832 tautomers	3	<0.1

This is in contrast to program packages that attempt accurate enumeration of tautomers and ligand protonation states under biological conditions. One such program is Schrödinger's pK_a prediction tool Epik [29, 30]. Epik purposely avoids energetically unfavorable forms (because they would create results for docking experiments that are undesired anyway), hence a much smaller number of generated tautomers can be expected. This is exactly what we found when we applied Epik to a small subset of 700 structures chosen to represent the tautomeric diversity in CSDB and compared the number of generated tautomers to the results of our approach. Such comparisons are therefore of limited relevance for the questions we tried to address in this study.

The price for our comprehensive approach is that we may, in some cases, tautomerically equate structures with each other that have such a high energy barrier for interconversion that they are in reality separate, stable compounds that do not interconvert even long-term. As already mentioned, to answer these questions quantitatively for individual compounds is beyond the means of current chemoinformatics since it requires careful analysis at the orbital level with consideration of the molecule's environment.

Another important aspect of tautomerism is that different tautomers of the same chemical compound vary in their pattern of double bonds, the form of specific functional groups taking part in the tautomerism, and the position of hydrogen atoms. This has an important consequence for bit-vector representations of molecular structures based on absence or presence of specific fragments and paths in the structure as commonly used for the calculation of Tanimoto-type similarity indices, which thus can be strongly affected by these kinds of structural changes. Therefore, any database searches based on Tanimoto similarities can be affected by tautomerism. Our extremely large tautomer set offered the ideal opportunity to analyze the magnitude of this effect quantitatively.

As part of the generation of the 680 million tautomer structure set, we calculated the Tanimoto similarity between each tautomer and the corresponding canonical tautomer (FICuS parent structure). Table 8 lists the distribution of calculated Tanimoto indices. The Tanimoto similarities were calculated using the PubChem fingerprints [28], which are a fragment-based bit-vector type representation of a chemical structure based on the CACTVS E_SCREEN bit vectors.

It stands to reason that calculating Tanimoto indices using fingerprints based on a different selection of fragments and paths, especially if these can be affected to a different degree by moving protons and double bonds in the context of tautomer enumeration, can be expected to yield different results. It was not possible to repeat this analysis with several different fingerprint types for the entire set of structures in CSDB. We did, however, conduct one, admittedly anectodal, comparison with one single

Table 8 Distribution of Tanimoto similarities in the entire set of tautomers (680 million structures) between the FICuS parent structure (canonical tautomer) and all derived tautomers]

Tanimoto index range	Count	%
>0.0–0.2	0	0.0
>0.2–0.3	6	<0.1
>0.3–0.4	6,580	<0.1
>0.4–0.5	369,331	<0.1
>0.5–0.6	6,304,436	0.9
>0.6–0.7	36,448,651	5.3
>0.7–0.8	111,954,384	16.4
>0.8–0.9	214,747,976	31.5
>0.9–1.0	310,725,465	45.6

The Tanimoto similarities were calculated using the PubChem fingerprints

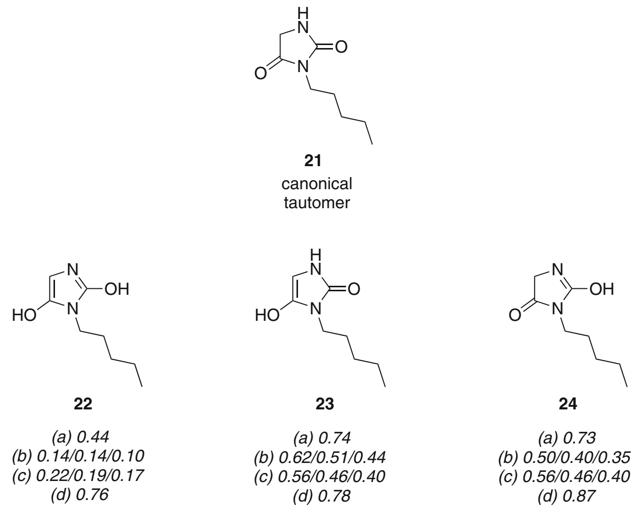


Fig. 8 Low Tanimoto similarity between different tautomers. Structure **21** is regarded as the canonical tautomer by CACTVS, structure **22**–**24** are formal tautomers generated from **21**. The italic numbers are the Tanimoto similarity indices between the canonical tautomer and respective tautomer structure calculated by **a** PubChem/CACTVS E_SCREEN fingerprints (881 bit fragment set), **b** extended connectivity fingerprints (FCFPs) with length of 2, 4, and 6 as implemented in Pipeline Pilot (1024 bit hashed), **c** functional class fingerprints (ECFPs) with lengths of 2, 4, and 6 as implemented in Pipeline Pilot (1024 bit hashed), and **d** MDL Public Keys as implemented in Pipeline Pilot

molecule (Fig. 8) for the following seven other fingerprinting methods: the functional class fingerprints (FCFP), the extended connectivity fingerprints (ECFP), each calculated for "lengths" (size of evaluated atom spheres around each heavy atom) 2, 4, and 6, and the MDL Public Keys as implemented by Pipeline Pilot [31–33].

Figure 8 shows the Tanimoto similarities between the canonical tautomer (**21**) and the alternative tautomeric forms (**22–24**) of the same chemical compound (NSC 68797). When compared by CACTVS/PubChem fingerprints (values (a)), Tanimoto similarity values all the way down to the 0.4 range were found. The corresponding values for the three (different-lengths) FCFP fingerprints and the three ECFP fingerprints, shown as sets of values (b) and (c), respectively, indicate that both the FCFP and ECFP fingerprints appear to be quite sensitive to tautomerism. The maximum similarity value found by FCFPs and ECFPs for tautomers **22–24** compared to the canonical tautomer was 0.62, while the minimum similarity value of 0.10 (tautomer **22**) was even lower than for the PubChem fingerprints. For the calculation of ECFPs, the number of connections, the number of bonds to non-hydrogen atoms, the atomic number, the atomic mass, the atomic charge, and the number of attached hydrogens are taken into account [34]. For the FCFPs, structural features like whether an atom is a hydrogen-bond donor, is positively ionizable, is negatively ionizable, is aromatic, or is a halogen are evaluated [34]. Most of these features changes if a different tautomer is used for the calculation.

On the other end, the MDL Public Keys turned out to be the fingerprints least sensitive to tautomerism. The lowest of the Tanimoto similarities for the structures in Fig. 8 (shown as values (c)) was 0.76. For the handful of structures we analyzed besides NSC 68797 (**21**) for this question the MDL Public Keys were usually less tautomerism-sensitive than the PubChem fingerprints. Both the PubChem fingerprints and the MDL Public Key fingerprints are calculated on the basis of predefined fragment sets. Without going to great lengths, our quick qualitative analysis of the fragments covered by the MDL Public Keys seemed to indicate fewer fragments with explicitly defined hydrogen atoms than for the PubChem fingerprints, which would explain their smaller sensitivity towards tautomers.

All these numbers are noteworthy, and perhaps even worrisome, if one ponders the following question: If we already may miss up to ~23% of matches by Tanimoto similarity at a cutoff of 0.8 due to tautomerism of one and the same compound (for PubChem fingerprints), how many more matches may be missed if we try to find truly just similar, i.e. not identical, compounds, using the same Tanimoto cut-off? To explore this question quantitatively is, however, beyond the scope of this paper. This finding begs the question if a tautomer-invariant form of the Tanimoto similarity (or, more accurately, of the bit vector representations used to calculate it) may be something that may be worthwhile developing. The MDL Public Keys seem to come closest to this among the fingerprints tested, though they certainly are not completely tautomer-invariant.

Analysis of stereochemistry

During the calculation of the FICuS parent structure set and the enumeration of the 680 million tautomers, we also analysed how stereochemistry is affected. We did this separately for E/Z stereochemistry of double bonds and chirality of atomic centers. In 72,556 cases among the 70.6 million canonical tautomers (FICuS parent structure set), explicitly defined stereochemistry on a double bond had been deleted in the way shown in Fig. 5. In 2,049,150 cases (2.9% of the FICuS parent structure set), we registered problematic stereochemistry on atomic centers. However, we did not apply the analogous treatment of removing the stereochemistry from these atomic centers, for the reasons discussed before. These numbers should be placed in the context of 8,079,330 of the FICuS parent structures (11.4%) being classified as possessing fully specified stereochemistry. During the generation of the 680 million tautomer set, we also tallied for how many canonical tautomer structures tautomers with potential stereocenters on atoms or bonds were generated. Both events occurred quite frequently: In 43,381,751 cases (61.4%), at least one tautomer was generated that had one or more double bonds being a potential E/Z stereocenter; potential atomic stereo centers were created for 30,818,806 of the canonical tautomers (43.6%).

Conclusions

According to the tautomerism definition used for the work described in this paper, tautomerism is not a rare occurrence in databases of truly existing compounds. Tautomerism was found to be possible for more than 2/3 of the unique structures in the CSDB. For nearly 5% of the 103.5 million original structure records did we find a case of either local or global tautomerism overlap. Projected onto the set of unique structures (by FICuS identifier), this still occurred in about 1.5% of the cases. Tautomeric overlap within each individual database in CSDB occurred on average for 0.3% of each database's entries, with values found as high as nearly 2% for some databases. When this analysis was extended to tautomeric overlap across all constituent databases in CSDB, the apparently more frequent occurrence of "tautomerism-critical" molecules across the 150+ individual databases caused the rate of overlap to jump to nearly 10%. In other words, unless one uses a tautomer-invariant approach, one has a nearly one-in-ten chance of missing a match when trying to match any one entry in CSDB with every other structure in our aggregated collection.

As discussed, the tautomerism definition (i.e. the ensemble of tautomeric transform rules) used in our tautomer-related work is rather comprehensive, certainly more so than in many other approaches and software used. Apart from very costly large-scale quantum-chemical calculations, it may take some ingenious and also not cheap experimental work to come to a verdict on which tautomerism definition best represents, at least in a statistical way, the practical situation encountered with databases of existing samples. In general, we believe it is important, from a chemoinformatics point of view, to have a tool for finding structures tautomerically linked to each other even if these tautomers may exist as separable compounds under certain conditions. As we have shown, a tautomerism analysis done right always should allow one to go back to the individual original structure (connectivity). The distribution of number of intra-database tautomer conflicts across the individual CSDB databases also argues against our tautomer definition being unreasonably broad.

We believe that these numbers also indicate that a more careful de-duplication of tautomeric multiples appears to be warranted for many existing databases, whereas some other small-molecule collections appear to be quite "clean" in this regard. Our analyses also seem to point to the necessity of considering the need for tautomer-invariant bit-vector structure representations and ensuing Tanimoto (and related) similarity calculations. All in all, tautomerism appears to be a topic that will be with the chemoinformatics and small-molecule database community for a while.

References

1. IUPAC Compendium of chemical terminology (electronic version). http://goldbook.iupac.org/T06252.html. Accessed Jan 26, 2010
2. Raczynska ED, Kosinska W, Osmialowski B, Gawinecki R (2005) Tautomeric equilibria in relation to pi-electron delocalization. Chem Rev 105(10):3561–3612
3. Pospisil P, Ballmer P, Scapozza L, Folkers G (2003) Tautomerism in computer-aided drug design. J Recept Signal Transduct Res 23(4):361–371
4. Trepalin SV, Skorenko AV, Balakin KV, Nasonov AF, Lang SA, Ivashchenko AA, Savchuk NP (2003) Advanced exact structure searching in large databases of chemical compounds. J Chem Inf Comput Sci 43(3):852–860
5. Milletti F, Storchi L, Sforna G, Cross S, Cruciani G (2009) Tautomer enumeration and stability prediction for virtual screening on large chemical databases. J Chem Inf Model 49(1):68–75
6. Oellien F, Cramer J, Beyer C, Ihlenfeldt W, Selzer PM (2006) The impact of tautomer forms on pharmacophore-based virtual screening. J Chem Inf Model 46(6):2342–2354
7. Smith M, Smith MB, March J (2007) March's advanced organic chemistry. Wiley, New Jersey
8. ChemNavigator Home Page. http://www.chemnavigator.com. Accessed 26 Jan 2010
9. The PubChem Project. http://pubchem.ncbi.nlm.nih.gov. Accessed 26 Jan 2010

10. Walker SB (2010) Personal communication
11. Ihlenfeldt W, Voigt JH, Bienfait B, Oellien F, Nicklaus MC (2002) Enhanced CACTVS browser of the open NCI database. J Chem Inf Comput Sci 42(1):46–57
12. Enhanced NCI Database Browser Webseite (2001) NCI/CADD Group, Frederick MD. http://cactus.nci.nih.gov/ncidb2. Accessed 26 Jan 2010
13. Blum LC, Reymond J (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 131(25):8732–8733
14. ChemNavigator iResearch^{TM} Library. http://www.chemnavigator.com/cnc/products/iRL.asp. Accessed 26 Jan 2010
15. NCI/CADD Chemical Identifier Resolver. http://cactus.nci.nih.gov/chemical/structure. Accessed 9 Jan 2010
16. NCI/CADD Chemoinformatics Tools and User Services. http://cactus.nci.nih.gov. Accessed 30 Jan 2010
17. PubChem FTP Server. ftp://ftp.ncbi.nlm.nih.gov/pubchem/Substance/CURRENT-Full. Accessed 28 Jan 2010
18. Ihlenfeldt WD, Takahashi Y, Abe H, Sasaki S (1994) Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. J Chem Inf Comput Sci 34(1):109–116
19. Xemistry chemoinformatics. http://xemistry.com. Accessed 26 Jan 2010
20. Sitzmann M, Filippov IV, Nicklaus MC (2008) Internet resources integrating many small-molecule databases. SAR & QSAR in Env Res 19(1):1–9
21. Ihlenfeldt WD, Gasteiger J (1994) Hash codes for the identification and classification of molecular structure elements. J Comput Chem 15(8):793–813
22. Leach AR, Bradshaw J, Green DVS, Hann MM, Delany JJ (1999) Implementation of a system for reagent selection and library enumeration, profiling, and design. J Chem Inf Comput Sci 39(6):1161–1172
23. 3-Penten-2-one, (e)- NIST Webbook page. http://webbook.nist.gov/cgi/cbook.cgi?ID=C3102338. Accessed 27 Jan 2010
24. (Z)-3-penten-2-one NIST Webbook page. http://webbook.nist.gov/cgi/cbook.cgi?ID=C3102327&Units=SI. Accessed 27 Jan 2010
25. Noack K, Jones RN (1961) Conformational equilibria in open-chain alpha beta unsaturated ketones. Can J Chem 39(11):2225–2235
26. Bokareva O, Bataev V, Godunov I (2009) Structures and conformational dynamics of monomethylated derivatives of acrolein: a quantum-chemical study. J Mol Struct THEOCHEM 913(1–3):254–264
27. He D, Li Z, Ma M, Huang J, Yang Y (2009) Study of extraction characteristics of HPMBP. 1. Tautomer and extraction characteristics. J Chem Eng Data 54(10):2944–2947
28. PubChem Substructure Fingerprint. ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. Accessed 29 Jan 2010
29. Epik. http://www.schrodinger.com/products/14/4/. Accessed 11 Mar 2010
30. Shelley JC, Cholleti A, Frye LL, Greenwood JR, Timlin MR, Uchimaya M (2007) Epik: a software program for pK_a prediction and protonation state generation for drug-like molecules. J Comput-Aided Mol Des 21(12):681
31. Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/. Accessed 11 Mar 2010
32. Rogers D, Brown RD, Hahn M (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. J Biomol Screen 10(7):682–686
33. Joseph LD, Burton AL, Douglas RH, James GN (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42(6):1273–1280
34. Hassan M, Brown R, Varma-O'Brien S, Rogers D (2006) Cheminformatics analysis and learning in a data pipelining environment. Mol Diversity 10(3):283–299