Tautomer Identification and Tautomer Structure Generation Based on the InChI Code

Torsten Thalheim,[†,‡] Armin Vollmer,[†] Ralf-Uwe Ebert,[†] Ralph Kühne,[†] and Gerrit Schüürmann*[,†,‡]

UFZ Department of Ecological Chemistry, Helmholtz Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany, and Institute for Organic Chemistry, Technical University Bergakademie Freiberg, Leipziger Strasse 29, 09596 Freiberg, Germany

Received March 27, 2010

An algorithm is introduced that enables a fast generation of all possible prototropic tautomers resulting from the mobile H atoms and associated heteroatoms as defined in the InChI code. The InChI-derived set of possible tautomers comprises (1,3)-shifts for open-chain molecules and (1,*n*)-shifts (with *n* being an odd number >3) for ring systems. In addition, our algorithm includes also, as extension to the InChI scope, those larger (1,*n*)-shifts that can be constructed from joining separate but conjugated InChI sequences of tautomer-active heteroatoms. The developed algorithm is described in detail, with all major steps illustrated through explicit examples. Application to ∼72 500 organic compounds taken from EINECS (European Inventory of Existing Commercial Chemical Substances) shows that around 11% of the substances occur in different heteroatom−prototropic tautomeric forms. Additional QSAR (quantitative structure−activity relationship) predictions of their soil sorption coefficient and water solubility reveal variations across tautomers up to more than two and 4 orders of magnitude, respectively. For a small subset of nine compounds, analysis of quantum chemically predicted tautomer energies supports the view that among all tautomers of a given compound, those restricted to H atom exchanges between heteroatoms usually include the thermodynamically most stable structures.

INTRODUCTION

Isomerism denotes the rearrangement of bonds within a given molecule. Tautomers are isomers that can be transformed to each other through chemical equilibrium reactions (typically with free energy changes below 25 kJ/mol). Prototropic tautomerism refers to bond changes accompanied with the intramolecular movement of hydrogen atoms. Prominent examples involving hydrogen exchange between carbon and heteroatoms include the keto−enol and imine−enamine equilibria. For a given compound, different tautomeric forms may differ in the number and type of certain functional groups, such as H bond donor and acceptor sites, which in turn result in tautomer-specific affinities for media, such as water and organic solvents.

Because a given tautomer represents a distinct chemical structure, the level of handling tautomers may have significant effects on the success rate in drug research and virtual screening as well as with regard to substructure search in large databases.[1−5] Typically, one tautomeric form is energetically preferred in a given medium. Then, any model to predict a property from the chemical structure will be incorrect, if an inadequate tautomer is submitted to the model. In other cases, the situation may be more complicated, and several different or even all possible tautomers of a particular compound may need to be taken into account, instead of considering a more or less arbitrary single structure. Pospisil et al.[6] and Martin[7] describe the impact of tautomerism on the quality of property prediction while using the appropriate tautomer form. In any case, the first step to address these problems is to identify all chemically reasonable tautomers.

For the latter, the typical approach is to employ rule-based transformations of a given initial structure to identify and enumerate all associated tautomers. Despite the flexibility and transparency of this approach, some new problems appear: many rules are accompanied with exceptions that in turn would require the development of additional rules for a fully automatized handling. For example, the rules of the Daylight SMIRKS notation introduced by Leach[8] were used for tautomer patterns by Oellien.[2] Here, 21 rules address particular occurrences of tautomerism. Of these, two rules identify (1,3)-shifts of H atoms of imines. This requires some exceptions for ring systems and for a possible terminal nitrogen atom. Another rule relates to an aromatic ring with two terminal heteroatoms. Without further specification, 1,3-benzenediol would wrongly match. In this latter case, an exception is defined, requiring one terminal atom to be nitrogen whereas the other atom could be either nitrogen or oxygen. Another rule-based algorithm is implemented by Trepalin et al.,[9] with each rule describing a certain type of tautomer pairs.

A further approach is given by Haranczyk and Gutowksi.[10] Here, the user has to identify all atomic sites and bonds of the molecule involved in tautomeric transformations (heavy atoms, mobile hydrogen atoms, respective bonds). The algorithm then excludes implausible structures based on a set of constraints, and predicts the most stable tautomer. Another kind of tautomer generation has been presented by Harańczyk et al.[11] Here, a carbon skeleton is deduced from the initial molecular structure. With a boolean vector (named fingerprint) the hydrogen atoms are assigned to their con-

* To whom correspondence should be addressed. Tel: +49-341-235-1262. Fax: +49-341-235-1785. E-mail: gerrit.schuurmann@ufz.de.
† UFZ Helmholtz Centre for Environmental Research.
‡ Technical University Bergakademie Freiberg.

10.1021/ci1001179
Published on Web 06/29/2010

necting atoms, and an enumeration process allows an easy generation of all possibilities. In addition, the fingerprint could be used very simply to detect symmetry. This method is the most similar to our approach.

Milletti et al.[12] use a simple (1,3)-shift scheme to create prototropic tautomers. From a given initial structure, all tautomers based on the shift scheme will be generated. The procedure is applied again to all newly generated structures. This will be repeated, until no new structures appear. As this recursive execution is very exhaustive, some strategies to avoid redundancy were introduced. This includes the use of InChI[13] to detect symmetric structures. The recursive calculation is skipped for structures that are generated for the second time. In addition, the energy difference and energy barrier to interconvert between both forms of an (1,3)-shift are analyzed during the generation. If the energy difference exceeds a threshold, the recursion is skipped too. Finally, a fragment- and pK_a-based evaluation is performed to predict the generated structure's stability.

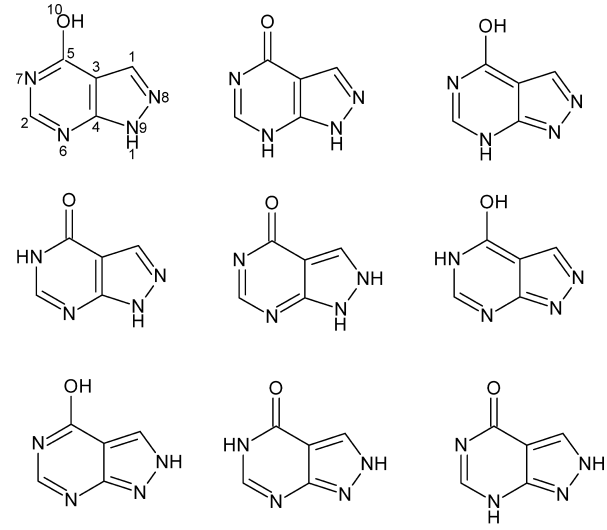
For the computerized generation of tautomers, several software tools have been developed.[10,14−18] TautTGen,[10] ConGENER,[14] and ACDChemSketch[15] are freely available; the source codes of TautTGen and ConGENER are provided via sourceforge.net,[10,14] and the algorithm has been published.[1,11] MN.TAUTOMER[16] is based on a set of eight rules[18] and can be tested with an online demo version. SPARC[17] offers an online application, but there is no explicit explanation of how the algorithm works.

ACDChemSketch is freeware for educational and private use, but the tautomer generation algorithm is not explained in detail in the freeware version. The tautomer is supposed to be independent from the initial structure because hydrogen atoms disposable to move will be detected automatically. However, this is restricted in some cases as can be illustrated with allopurinol ($C_5H_4N_4O$, 1*H*-pyrazolo[3,4-*d*]pyrimidine-4-ol, IUPAC name 3,5,7,8-tetrazabicyclo[4.3.0]nona-3,5,9-triene-2-one), a drug used to treat hyperuricema (chronic gout) through inhibition of xanthine oxidase and the associated generation of uric acid.

Scheme 1 shows the nine tautomers of allopurinol. With ACDChemSketch, only the tautomers with O═, except the structure in the center of Scheme 1 (with both H atoms at the 5-ring N atoms), will be generated. Consequently, if one of these four structures is selected as input, the remaining three are generated, resulting in a total of four tautomers identified. In all other cases, the ACDChemSketch algorithm yields five tautomers: The four structures just mentioned and the input structure. The reasoning behind this algorithm might be a pragmatic approach to avoid energetically unreasonable tautomers, which however is neither documented nor demonstrated as such.

In recent years, the great potential of the International Chemical Identifier, InChI,[13] to store and handle chemical information on the connection table level in large chemical databases has been recognized. In addition to line notations, such as the SMILES code,[19] to define the molecular topology and bond types, the InChI code is becoming very popular and can also be used as unique identifier, such as the CAS number.[20] Even though InChI is proprietary, it is freely available from the intellectual right holders, including computerized versions even with the source code. Interestingly, InChI contains pertinent information about prototropic

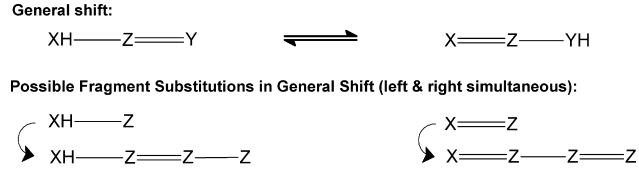
Scheme 1. Nine Tautomers of Allopurinol and the Associated InChI Code Specifying the Mobile H Atoms (the Latter of Which Are Given in Bold)[a]



InChI=1/C5H4N4O/c10-5-3-1-8-9-4(3)6-2-7-5/h1-2H,**(H2,6,7,8,9,10)**

[a] The InChI atom numbering (which differs from the IUPAC numbering) is specified in the top left tautomer. As can be seen from the InChI code (bottom), the "H2" notation indicates the presence of two mobile H atoms, which can be attached to any two of the five heteroatoms #6−10 covering four nitrogens (#6−9) and one oxygen (#10).

Scheme 2. Prototropic Tautomerism Because of (1,3)-Shifts of H Atoms between Heteroatoms[22] (Upper Part) and H Atom Movement Extended to (1,*n*)-Shifts, with *n* being Odd and Larger than 3, upon Insertion of Conjugated Fragments (Lower Part)[a]



[a] The InChI definition of mobile H atoms and associated heteroatom acceptor sites covers all respective (1,3)-shifts for open-chain structures, and larger (1,*n*)-shifts across conjugated ring systems. In addition to that, our algorithm includes also (1,*m*)-shifts (with odd *m* > odd InChI *n*) constructed from joining separate but conjugated (1,*n*)-shifts.

tautomers with regard to mobile hydrogen atoms attached to heteroatoms. The reasoning behind focusing on this subset of tautomers is that for a given chemical structure, the thermodynamically most stable tautomer is often among those that are related to each other through hydrogen exchange between heteroatoms accompanied by respective exchanges of single and double bonds.

The aim of our study is to provide an algorithm to construct all prototropic tautomers that can be generated from the mobile H atoms and associated heteroatom acceptor sites specified through the InChI code. The respective tautomeric equilibrium is outlined in Scheme 2, specifying both the types of heteroatoms considered and the associated change between single and double bonds upon movement of a given H atom between different heteroatoms of the molecule. In this way, the algorithm enables a fast identification of stable tautomers and thus provides pertinent information for both structure

and substructure searches, as well as for structure-based predictions of compound properties, such as partition coefficients and acid−base equilibria. As shown recently for the example of N-hydroxyl amidines,[21] H atom shifts between heteroatoms typically preserve the structure stability better than hydrogen shifts involving carbon atoms.

INCHI CODE INFORMATION ON TAUTOMERS

The IUPAC International Chemical Identifier (InChI)[22] was introduced in 2000. InChI serves as unique identifier for chemical substances, and with its unambiguous notation allows for an easy comparison of chemical structures. Thus, it is applied in various free and commercial databases, such as the U.S. National Institute of Standards and Technology[23] or the Kyoto Encyclopedia of Genes and Genomes.[24]

The InChI code covers six major layers, of which the last two (fixed-H layer and reconnected layer for metals) are optional. The first major layer provides the main features of the molecular structure and consists of three sublayers: chemical formula, connection table without specification of bond types and without H atoms, and H atoms covering both hydrogens attached to specific sites (immobile H atoms) and so-called mobile H atoms together with lists of heteroatoms as possible (but at this stage not fixed) attachment sites. These InChI-defined mobile H atoms and their associated heteroatom acceptor sites provide the key input information for our algorithm to generate, from a given molecular structure, all respective tautomeric forms. The remaining major layers 2−4 of InChI are the charge layer, the stereochemical layer, and the isotopic layer, all of which are not needed for our algorithm.

The fifth major layer allocates the mobile H atoms to particular heteroatoms, thus yielding a uniquely defined molecular structure. Our algorithm uses this layer to remove, after initial generation of all possible tautomers according to the InChI-defined mobile H atoms and their heteroatom acceptor sites, all duplicates, thus yielding the final set of structurally different prototropic tautomers (see below).

Note further that the InChI code can be specified in two different forms: The InChI string as also given in our examples (see Schemes 1, 5, and 6), and its associated hash code preferred for machine reading, the InChI key. Use of the latter speeds up the tautomer comparison significantly, and it is used in our algorithm.

MATERIALS AND METHODS

Algorithm. The algorithm has been coded in C++ and developed as module of our in-house software ChemProp.[25] It runs on Windows PCs under standard configurations.

After input of the molecular structure in a standard format, such as the InChI code, the SMILES code, the MOLfile, or the SMDfile format, ChemProp converts the structural information into an internal format. In case the input format was not InChI, the InChI-code information is generated internally. Subsequently, the algorithm calculates the number of all different prototropic tautomers available from the InChI-defined mobile H atoms and associated tautomer-active heteroatoms. Finally, all respective tautomers are generated and stored in a fully automatized manner, and thus are made available for later inspection or further separate use.

Tautomer Energy. For a small set of tautomers, molecular energies in terms of heats of formation at 25 °C, ΔH_f, have been calculated employing the semiempirical quantum chemical AM1[26] method as implemented in MOPAC.[27] To this end, initial 3D geometries have been generated within ChemProp, and the final geometry optimization has been performed at the AM1 level employing MOPAC.

QSAR Calculations. To test the variation of QSAR predictions across tautomers, logarithmic values of the soil sorption coefficient normalized to organic carbon, log K_{oc}, and of the water solubility at 25 °C, log S_w, have been calculated with literature methods[28,29] as implemented in ChemProp.

Molecular Structures. For the development and application of the tautomer-generation algorithm, ∼72 500 organic molecules have been retrieved from the list of European Inventory of Existing Commercial Chemical Substances (EINECS)[30,31] and stored in a ChemProp database.

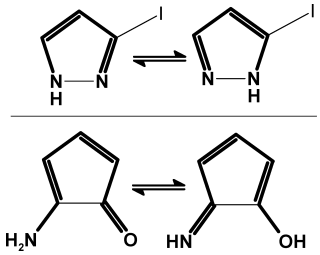
RESULTS AND DISCUSSION

Tautomer Identification. To identify prototropic heteroatom-confined tautomerism as outlined in Scheme 2, it is just necessary to compare the main InChI layer and the charge layer. The other layers will be disregarded at this stage. Since InChI generates a canonical atom enumeration (for non-H atoms) by a modified Morgan algorithm,[32] different molecules with the same main layer are tautomers of each other. Thus, restriction to the main layer yields a generic code for all tautomers of a particular compound. Specification of a particular form then requires extension to the fifth main layer (see above). In other words, comparison of two structures in terms of being tautomeric forms of a common general structure can very easily be achieved by simple string operations. The real challenge is to generate all possible tautomer forms for a given structure.

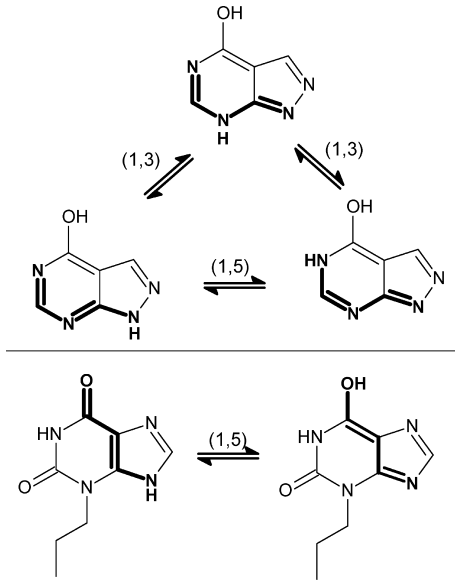
Tautomer Generation. Since the InChI code already specifies the mobile H atoms, the task left is a procedure to automatically generate, from any of the possible tautomeric forms as starting structure, all respective tautomers. To this end, a brute-force approach can be applied: From the structure of a given compound, the InChI code is generated, and from the resulting mobile H sublayer the specified number of H atoms simply has to be attached to all possible heteroatom positions in a combinatorial manner in accord with chemical bonding rules.

In the next step, a completion process for each generated structure is required to obtain a valid configuration of valences and bonds. Within InChI, heteroatoms acting as possible H atom acceptors may have single bonds only (sp^3 hybridization) or are involved in double bonds (sp^2 hybridization). Here, attachment or detachment of hydrogen is accompanied by a respective change in the hybridization. All other atoms not involved in the H atom movement must retain their initial hybridization in all tautomers. Correspondingly, the types of the bonds not involved in tautomerism remain the same for all tautomers. Moreover, each single bond converted to a double bond upon tautomerism is balanced by a double bond becoming a single bond. It follows that the number of double bonds of a given compound is also constant for all tautomers. Because this number of double bonds is not directly apparent from the

Scheme 3. InChI-Based Prototropic Tautomeric Equilibria Involving an (1,5)-Shift (Top) and an (1,7)-Shift (Bottom) Across Conjugated Ring Structures As Examples of (1,*n*)-shifts with *n* Being Odd and Larger than 3



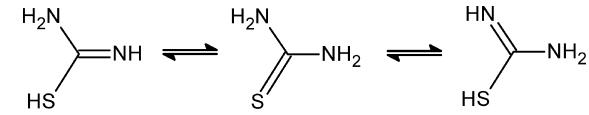
Scheme 4. Examples of an (1,5)-Shift Across Two Fused Rings and Its Equivalent Representation As Consecutive Sequence of Two (1,3)-Shifts (Top) and of a Nonreducible (1,5)-Shift That Is Outside the InChI Scope but Addressed through Our Algorithm (Bottom)



InChI Code, our algorithm takes this value from the initial structure used to generate all respective tautomers, and adds the appropriate number of double bonds to each of the tautomers generated.

In terms of hydrogen shifts, the mobile H atoms defined through InChI refer primarily to (1,3)-shifts of open-chain molecules (Scheme 2), and for ring systems also to (1,*n*)-shifts with *n* being an odd number larger than 3. Two examples of the latter are given in Scheme 3, illustrating an (1,5)-shift (top) and an (1,7)-shift (bottom) through respective participation of ring atoms. Because InChI does not consider (1,*n*)-shifts (with $n = 5, 7, ...$) across two (or more) rings of fused systems (except those that can be reduced to equivalent sequences of (1,3)-shifts), respectively defined mobile H atoms are missing as basis for generating the associated tautomeric forms. However, our algorithm provides the following extension to the InChI set of possible tautomers: If two InChI-based H atom shifts can be combined to a larger shift across an additional conjugated fragment, this larger shift will be detected and taken into account. Scheme 4 contains an example of an (1,5)-shift across two fused rings and its equivalent representation as consecutive sequence of two (1,3)-shifts (top), and an example of a nonreducible (1,5)-shift that is outside the InChI scope, but addressed through our algorithm (bottom). According to present experience, our algorithm yields approximately 1/3 more tautomers than the original InChI approach in those cases where the structure contains conjugated fragments linking otherwise isolated H atom shifts.

Scheme 5. Symmetric Heteroatom Acceptors (Nitrogen Atoms) for Tautomeric H Shifts in Carbamimidothioic Acid[a]



InChI=1/CH4N2S/c2-1(3)4/**(H4,2,3,4)**

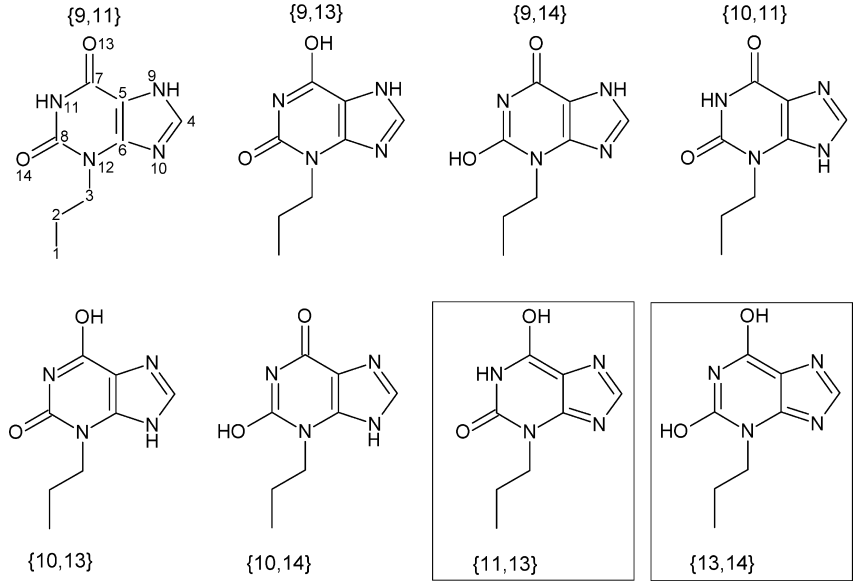
[a] Exchange of two H atoms between the nitrogen groups yields an identical structure, only mirrored at the S−C bond.

For a further illustration of our algorithm, we come back to the drug allopurinol (Scheme 1, with the InChI code given at the bottom). According to the canonical formula represented in InChI as C5H4N4O, the numbering is done by starting with all carbon atoms followed by nitrogen atoms and the oxygen atom. The next sublayer c10-5-3-1-8-9-4(3)6-2-7-5 sketches the molecule skeleton similar to a structural formula by referring to carbon atoms 1−5, nitrogen atoms 6−9, and oxygen atom 10. This is shown in all tautomer skeletons in Scheme 1. The sublayer /h finally assigns the H atoms: Two of the four H atoms are fixed at carbon atoms 1 and 2 at the respective chain nodes by the string 1-2H, and the mobile hydrogen atoms are left in (H2,6,7,8,9,10). H2 denotes two H atoms, attached to two of five possible positions offered by the four nitrogens (#6,7,8,9) and the one oxygen (#10). Thus, $\binom{5}{2} = 10$ possible arrangements have to be checked. Nine of them yield the valid structures shown in Scheme 1. The only case without success is the selection of the nitrogen atoms #6 and 7 that are separated by carbon atom #2. Parallel H attachment to both of these N atoms is not compatible with a closed-shell conjugated system. The technical procedure to extract InChI information and to find valid structures will be explained in the next section in more detail.

In the final step of the tautomer generation, the resultant valid structures need to be compared to remove duplicates. Taking carbamimidothioic acid (Scheme 5) as example, it is seen that two of the initially generated three tautomers (left and right) are identical, just mirrored at the S−C bond. Other examples are carbonyls. Interestingly, InChI marks COOH groups to have a mobile hydrogen atom, although the resultant (formally correct) tautomers with H at the one or other O atom cannot be distinguished. In principle, such duplicates could be avoided by a respective algorithm in the combinatorial step. However, this would not result in a significant gain of performance, so the suggested procedure is to compare the results, and to simply remove identical structures.

The latter requires a method for comparing molecular structures, for which there are several techniques available. One approach, similar to the Milletti[12] method, is to generate the InChI keys of the structures, this time with the option *mobile H perception* deactivated. The result is an additional layer with defined attachments of all H atoms, including the tautomer-mobile hydrogens. Using the key enables a simple removal of duplicates: Every key must appear only once, and thus duplicate keys and their related structures can be deleted easily.

Scheme 6. Eight Tautomers of Enprofylline[a]



InChI=1/C8H10N4O2/c1-2-3-12-6-5(9-4-10-6)7(13)11-8(12)14/h4H,2-3H2,1H3,(H,9,10)(H,11,13,14)

[a] The InChI code given in the bottom refers to the first six tautomers, specifying the mobile H atoms and associated heteroatom acceptor sites. The additional two tautomers, shown as framed structures at the bottom right, are not covered by the InChI code given because of its separate InChI shift options (H,9,10) and (H,11,13,14). However, they become included when employing the combined sequence (H2, 9, 10, 11, 13, 14) as is done with our tautomer-generation algorithm. The numbers in brackets specify the individual tautomers in terms of the heteroatoms (InChI numbering) selected as H atom attachment sites.

Algorithm for Generating Individual Tautomers. The algorithm introduced here involves three steps: First, the heteroatoms available for hydrogen attachment are identified, and the mobile H atoms as defined through InChI are subdivided into tautomer-active and tautomer-passive subsets. The latter subset comprises those H atoms that despite their principally mobile character remain fixed because of valence rules (see below). At this stage, carbon−carbon bonds involved in tautomerism are not yet identified.

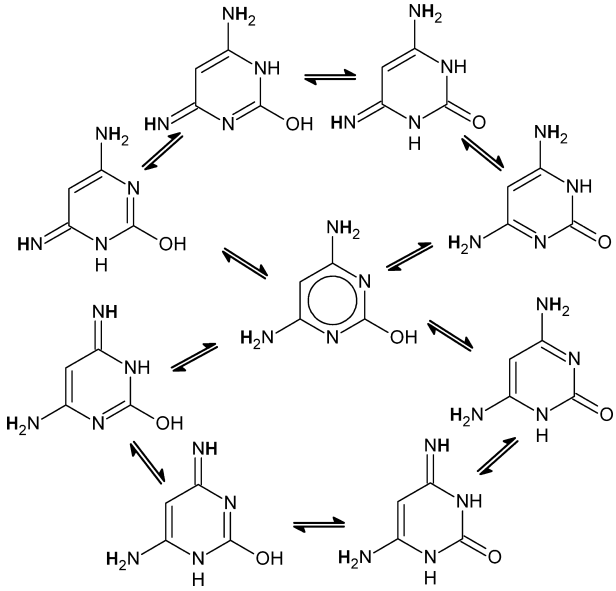
In the next step, all atoms and bonds not participating in tautomeric rearrangements will be removed, yielding a tautomer skeleton with single bonds only. Then, the attachment sites for the hydrogen atoms are specified in a combinatorial manner. Third, the algorithm tries to define double bonds along the atom−atom connections of this skeleton with the attached hydrogens in a conjugated manner in all possible constellations. If this is not possible, the skeleton is rejected and no longer tautomer-relevant. Otherwise, the accordingly generated molecular structures are checked for duplicates that are deleted as well, yielding the final set of prototropic tautomers.

Step 1: Identification of Heteroatom H Acceptors and of Tautomer-Active H Atoms. All H atom-accepting heteroatoms are directly available from the respective InChI code fragment. However, the InChI information on mobile H atoms needs to be modified for the purpose of tautomer generation.

First, multiple InChI sequences of mobile hydrogens are joined into a single sequence. Taking the InChI code of enprofylline as example (Scheme 6, bottom), the notation (H,9,10)(H,11,13,14) indicates the following two sequences of different mobile H atoms: One mobile H atom can be attached at sites 9 and 10 (that means at one of the two 5-ring nitrogens) and another mobile H atom at one of the sites 11, 13, and 14 (6-ring heteroatoms except the propyl-substituted nitrogen). The reason for that splitting is to avoid implausible structures. However, this splitting results in an a priori exclusion of two tautomers with no mobile H atom attached to the 5-ring nitrogens (see the two framed structures at the bottom right of Scheme 6). Accordingly, it appears more satisfactory to initially allow for the generation of an upper bound of tautomer candidates and then exclude valence-incorrect settings as well as duplicates. Thus, in case of enprofylline our algorithm converts the separate InChI sequences (H,9,10)(H,11,13,14) into the combined sequence (H2,9,10,11,13,14), now indicating that there are two, formally equivalent, mobile H atoms that can be attached to any two of the five heteroatom sites available. In this way, also tautomer candidates with both H atoms attached at the 5-ring or at the 6-ring are taken into account. The subsequent valence check then reveals that the option with both H atoms at the two 5-ring nitrogens does in fact not lead to valid structures and is then excluded from further consideration.

Second, the mobile H atoms are checked for multiple occurrences at single sites. While all mobile H atoms identified through InChI are mobile *in principle*, the ones attached to the same heteroatom are not mobile at the same time. Moreover, if one of several mobile H atoms at a given heteroatom has been shifted to generate a new tautomer, then separate consideration of the remaining H atoms would yield identical results (duplicates) and thus no further new tautomer. In the algorithm, this issue is addressed as follows: Initially, the InChI-defined mobile H atoms are considered as potentially mobile. Then, each heteroatom acceptor site is checked for free valences in the first InChI sublayer, where all bonds are still single bonds and where the (potentially) mobile H atoms are not yet included. Subsequently, H atoms are attached to heteroatoms with free valences until the latter

Scheme 7. All Formally Valid Heteroatom−Prototropic Tautomers of 4,6-Diaminopyrimidine-2-ol, Covering the Unique Tautomer in the Center and Two Sets of Four Distinct Tautomers (Top Four and Bottom Four Structures Arranged in Half Circles) that are Duplicates of Each Other Because of Molecular Symmetry[a]



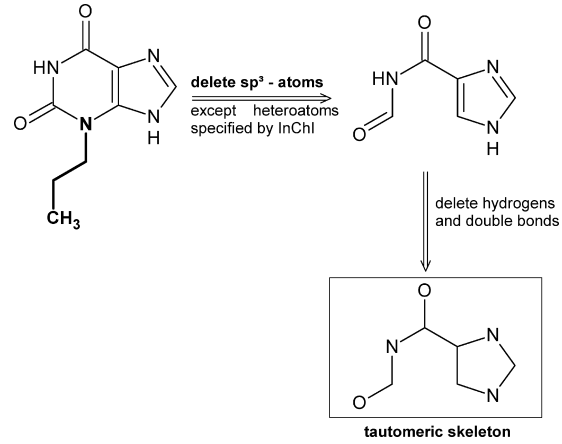
[a] Among these initially generated nine tautomers, the four duplicates (say, the ones located in the bottom half circle) are removed in step 4 of our algorithm (see text). Two of the InChI-mobile H atoms are tautomer-passive, because multiple H atoms at heteroatoms cannot be moved at the same time and because their separate movement would yield tautomers already generated otherwise (see step 1 as described in the text). The principally mobile but tautomer-passive H atoms are printed in bold.

have gone down to one. The remaining initially mobile H atoms are now classified as tautomer-passive H atoms, and no longer belong to the subset of tautomer-relevant H atoms. Correspondingly, the ones successfully attached to any of the heteroatoms are classified as tautomer-active H atoms, and thus form the final subset of mobile H atoms used for generating individual tautomers.

The approach is illustrated with 4,6-diaminopyrimidin-2-ol in Scheme 7. As can be seen, all nine initially generated tautomers, including four duplicates because of molecular symmetry (all four bottom tautomers are duplicates of located top counterparts, the only unique tautomer being the structure in the center of Scheme 7), contain at least one H atom at each of the two nitrogens. These principally mobile but tautomer-passive H atoms are printed in bold in the scheme. Subsequent exclusion of duplicates (as described in step 4 below) yields five different tautomers, covering the central structure, as well as the four structures of the four tautomers located in the upper half circle of Scheme 7.

Step 2: Tautomeric Skeleton Generation. Scheme 8 illustrates the algorithm to obtain the tautomeric skeleton with enprofylline taken from Scheme 6 (now using tautomer {10,11} as starting structure). Because sp^3-hybridized atoms except H-acceptor heteroatoms cannot participate in tautomeric rearrangements, respective subunits are identified and removed from the initial molecular structure. In case of enprofylline, this concerns the *N*-propyl subunit shown in bold in Scheme 8. Then, all H atoms and all atoms with triple bonds are removed. The latter is done because the InChI-generated mobile H atoms do not refer to tautomeric rearrangements involving triple-bonded atoms. After subsequent removal of all bonds not connecting two atoms, the remaining double bonds are counted because their number remains constant across all prototropic tautomers (see above) and then converted to single bonds, thus generating the maximum degree of free valences of all tautomer-relevant heteroatoms. The resultant tautomeric skeleton is now ready for tautomer generation. Note, however, that this skeleton is not suited for detecting ring−chain tautomerism, the latter of which is already outside the respective InChI scope (as defined through the InChI mobile H atoms and associated heteroatom acceptor sites).

Scheme 8. Construction of the Tautomeric Skeleton of Enprofylline, Employing the Structure with Mobile H Atoms Attached at Nitrogen #10 and #11 (Tautomer {10,11} in Scheme 6) as Starting Structure[a]

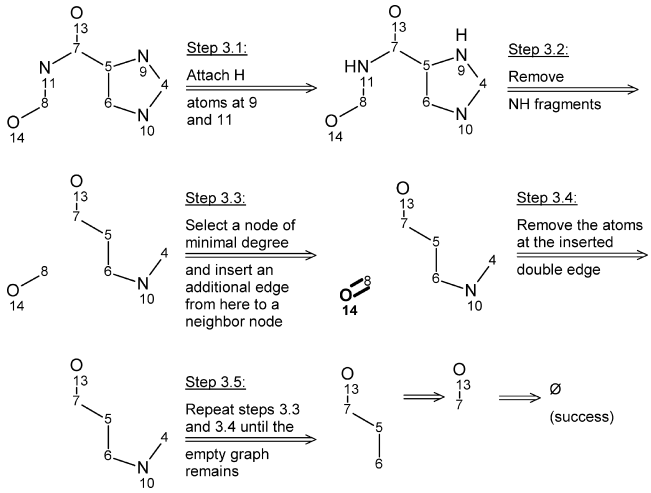


[a] Initial deletion of the tautomer-passive *N*-propyl unit is followed by subsequent removal of H atoms and double bonds to yield the final tautomeric skeleton. The latter is used as starting point to generate all tautomers as described in the text.

Step 3: Combinatorial Assignment of Tautomer-Active H Atoms and Double Bonds. First, the number of possible attachments of tautomer-active H atoms to heteroatom accepting sites is calculated. Then, each respective combinatorial subset (each particular setting of H attachments) is mapped onto the tautomeric skeleton, and for each such bonding pattern the relevant number of double bonds is introduced to meet the valence rules. In graph theoretical terms, this means that edges will be added to a simple graph, until the accordingly modified graph consists of a maximal number of double edges.

More precisely, the algorithm proceeds as follows: After initial attachment of the tautomer-active H atoms to particular heteroatoms, these latter have already reached saturation of their valences and have been identified as such and, temporarily, removed from the structural skeleton. Then, the atom of the remaining skeleton with the fewest unsaturated neighbors, called minimal node degree in Scheme 9, is selected (randomly in case of more than one equivalent option), and a double bond is introduced between this atom and an unsaturated neighbor. Through this step, the valences of the double-bonded atoms become saturated, and both bonding partners are, again temporarily, removed. This procedure is repeated step by step until the required number of double bonds (that had been calculated previously, see above) has been introduced. At this point in time, all valences of all atoms have been saturated, and the final bonding pattern of the particular tautomer is fixed. Now, the complete tautomeric structure is built through joining the initially

Scheme 9. Workflow to Generate from the Tautomeric Skeleton a Particular Completely Defined Tautomeric Structure, Taking Enprofylline As Example (Step 3 in the Algorithm, See Text)[a]



[a] During the stepwise procedure, two H atoms are attached at two particular heteroatom acceptor sites (here nitrogens #9 and #11), and four double bonds are generated before their (temporary) removal together with respective bonding partners. The result corresponds to the top-left tautomer {9,11} of Scheme 6, which is obtained through building the structure from all temporarily removed substructures according to the selected path.

removed tautomer-passive substructures (see step 2 above) and the tautomer-active structural units just described.

During the stepwise procedure of removing atom pairs connected through double bonds, some atoms, or nodes in graph theoretical notation, may become isolated. A node is isolated, when it is not saturated yet, but none of the neighbor atoms still contains free valences. In this case, the last steps will be revoked until an atom is reached that has more than one unsaturated neighbor. Then, for this atom a double-bond partner is selected that is different from the previous unsuccessful path, and the stepwise procedure continues as described above. If, however, no such alternative pathway can be found when going backward from an isolated node, the particular allocation of H atoms to heteroatom acceptors is invalid, and the respective combinatorial subset is rejected.

Step 4: Duplicate Removal. There are several straightforward approaches to achieve this task. A simple but effective opportunity is to compare the InChI keys of the resulting tautomers: After the creation of all valid subsets, the codes including the fixed fixed-H layer will be generated. In the original InChI software, this option is called "without the mobile H perception". This provides a means to detect identical solutions by hash code comparison. Duplicates detected in this way are then removed, resulting in the final set of different prototropic tautomers. Coming back to 4,6-diaminopyrimidin-2-ol in Scheme 7, the four tautomers in the lower part are duplicates of the four structures in the upper part, which is detected and accounted for at this step 4 of our algorithm.

Example. The procedure is illustrated further with the already discussed drug enprofylline. From the InChI code (Scheme 6, bottom), five H-accepting heteroatom sites (N atoms #9−11, and O atoms #13−14) are obtained, identifying the propyl-substituted nitrogen (atom #12) as tautomer-passive. In this case, the InChI number of mobile H atoms equals the number of tautomer-active H atoms, which is two. Generation of the tautomeric skeleton (step 2) through removal of tautomer-passive sp^3 atoms (here, *N*-propyl unit) has been illustrated in Scheme 8. Scheme 9 now shows the stepwise construction of the tautomer {9,11} with the two mobile H atoms attached at nitrogens #9 and 11, which forms a particular combinatorial subset of the $\binom{5}{2} = 10$ different settings of attaching the two mobile H atoms to any two of the five heteroatom acceptors (step 3). After initial attachment of the hydrogens at the nitrogens #9 and 11 (first substep of step 3 = step 3.1), these nitrogens have reached their valence saturation and are removed (step 3.2). Subsequently, double bonds (double edges) are introduced in a stepwise manner such that no isolated nodes arise (steps 3.3, 3.4, 3.5). For each of the 10 initial allocations of the mobile H atoms to tautomer-active heteroatoms, the procedure attempts to assign the required number of double bonds (edges), in this case four double bonds, to the resulting molecular fragments. However, the attachment settings {9,10} (H atoms attached to both 5-ring nitrogens) and {11,14} are rejected because of isolated nodes (node #4 resulting from {9,10}, and node #8 resulting from {11,14}). In step 4, the remaining eight valid settings are compared via the fixed-H layer layer of InChI. Because there is no duplicate structure detected, these eight structures differ from each other and form the final set of valid tautomers, already shown in Scheme 6.

Application to a Large Compound Set. To demonstrate the potential relevance of considering tautomers, our algorithm has been applied to ~72 500 organic molecules of the European Inventory of Existing Commercial Chemical Substances (EINECS).[30,31] In 11% of these structures, tautomerism based on H atom shifts between heteroatom occurs. For most of the compounds, the number of different tautomers does not exceed 20, and less than 40 substances have more than 500 tautomers. In total, the algorithm took about 25 min on an Intel Core II 2.4 GHz CPU to generate the tautomers. Depending on the available hardware, there may be a considerable amount of additional time to write the results on the hard disk, which is independent of the algorithm.

QSAR Variation Across Tautomers. The variation of QSAR predictions across tautomers has been examined through application of respective models for predicting the soil sorption coefficient normalized to organic carbon, K_{oc},[28] and water solubility (in mol/L at 25 °C), S_w,[29] of the EINECS compounds. Initial exclusion of compounds with an estimated number of tautomers larger than 2500 resulted in ~7900 tautomer sets with a total number of approximately 42 000 individual structures. For these, log K_{oc} and log S_w have been predicted using our in-house software ChemProp.[25] For this screening-level analysis, we have not considered the chemical domain,[33] keeping in mind that the latter becomes important when evaluating the confidence of QSAR predictions. Moreover, we did not exclude any of the generated tautomers on the basis of energetic arguments. Thus, the QSAR model[28] for predicting log K_{oc} could be formally applied to 7466 tautomer sets (comprising 7466 different compounds with a total of 39434 individual chemical structures), and the suite of methods[29] for predicting log S_w to 7774 tautomer sets (covering 41607 individual chemical structures).

The results clearly demonstrate the importance of tautomer consideration. The standard deviation of log K_{oc} across tautomers was almost 0.4, with individual K_{oc} differences

Table 1. Variation of QSAR Predictions for log K_{oc} and log S_w Across Tautomers and Associated Ranges of the Tautomer Energies[a]

compound name	chemical formula	no. InChI-based (heteratom-confined) tautomers	predicted property range: Δ (log K_{oc})	predicted property range: Δ (log S_w) [mol/L]	AM1 energy range of InChI-based tautomers: Δ (ΔH_f) [kJ/mol]	total no. tautomers	AM1 energy range of all tautomers: Δ (ΔH_f) [kJ/mol]
enprofylline	$C_8H_{10}N_4O_2$	8	1.11	0.38	118.28	14	192.34
primisulfron	$C_{14}H_{10}F_4N_4O_7S$	5	1.85	0.71	67.57	11	87.62
tribenuron	$C_{14}H_{15}N_5O_6S$	2	1.01	0.25	42.73	12	116.42
sulfalene	$C_{11}H_{12}N_4O_3S$	2	0.61	0.72	55.04	5	137.12
nitrofurantoin	$C_8H_6N_4O_5$	3	0.61	0.05	87.22	5	119.93
dimethoate	$C_5H_{12}NO_3PS_2$	2	0.24	0.32	44.48	3	44.77
icosanamide	$C_{20}H_{41}NO$	2	0.24	0.1	54.31	3	66.00
timiperone	$C_{22}H_{24}FN_3OS$	2	0.71	0.74	9.54	4	38.3
tromantadine	$C_{16}H_{28}N_2O_2$	2	0.24	0.32	61.04	3	87.09

[a] log K_{oc} (soil sorption coefficient normalized to organic carbon) and log S_w (water solubility in mol/L at 25 °C) have been predicted for all tautomers of the nine compounds with QSAR models taken from literature[28,29] as implemented in ChemProp.[25] The InChI-based tautomers confined to H atom shifts across heteroatoms were generated by our newly developed algorithm, and the additional tautomers involving sp^2 carbons in H atom shifts were constructed manually, resulting in a correspondingly larger total number of tautomers as specified in the second last column. The calculated heats of formation at 25 °C, ΔH_f [kJ/mol], were obtained from AM1[26] calculations including geometry optimization with the MOPAC software.[27]

Table 2. Variation in log K_{oc} and log S_w [mol/L] of the Tautomer-Active Subset of Organic EINECS Compounds in Terms of Value Ranges and Associated Numbers of Compounds, as Well as Numbers of Tautomers Per Compound[a]

number of tautomers per compound	no. of compounds for a given property range $x = \Delta$ (log K_{oc}): $x \le 0.5$	$0.5 < x \le 1$	$1 < x \le 1.5$	$1.5 < x \le 2$	$x > 2$	no. of compounds for a given property range $x = \Delta$ (log S_w) [mol/L]: $y \le 1$	$1 < y \le 2$	$2 < y \le 3$	$3 < y \le 4$	$y > 4$
≤5	5198	1044	326	31	21	6237	439	76	23	33
6−50	122	359	207	99	32	524	299	61	30	24
51−500	0	2	8	3	2	5	5	2	4	0
>500	0	0	0	0	12	0	8	1	3	0

[a] log K_{oc} (soil sorption coefficient normalized to organic carbon) and log S_w (water solubility at 25 °C in mol/L) have been predicted for all tautomers of the nine compounds with QSAR models taken from literature[28,29] as implemented in ChemProp.[25] The prototropic tautomers generated result from the mobile H atoms and associated heteroatom acceptor sites as defined in the InChI code. Omitting compounds with estimated numbers of tautomers larger than 2500 (see text); log K_{oc} was predicted for 7466 compounds across 39 434 individual tautomers, and log S_w for 7774 compounds across 41 607 individual tautomers.

between the tautomers of a given compound up to 4 orders of magnitude. For the log S_w prediction, the standard deviation across tautomers was 0.75, with maximum differences in S_w of more than nine logarithmic units.

Table 1 shows the variations in predicted log K_{oc} and log S_w for enprofylline, for primisulfron with five distinct tautomers, nitrofurantoin with three distinct tautomers, and for six further compounds with only two distinct tautomers. In addition, the molecular energy variation across tautomers is quantified through calculated heats of formation at 25 °C, ΔH_f, employing the semiempirical quantum chemical AM1[26] model including geometry optimization as implemented in MOPAC.[27] As can be seen from the table, the tautomer variation in log K_{oc} is not correlated with the one in log S_w, reflecting corresponding differences in the dependencies of both properties on molecular structure. Similarly, larger differences in log K_{oc} or log S_w are not necessarily associated with larger differences in ΔH_f.

These finding are illustrated with the calculated values obtained for enprofylline, primisulfron, sulfalene, and timiperone: Enprofylline yields a log K_{oc} variation across its eight tautomers of 1.11 as opposed to a corresponding log S_w variation of 0.38, accompanied by a ΔH_f variation of 118.3 kJ/mol. With primisulfron, both the log K_{oc} and log S_w variation are significantly increased, while the variation in ΔH_f is reduced by a factor of 2. Sulfalene and timiperone show similar variations in log K_{oc} and log S_w of about 0.6−0.7, but substantial differences in the associated ΔH_f variations (82.1 kJ/mol vs 29.3 kJ/mol; see Table 1).

Coming back to the EINECS subset of compounds with tautomerism through H atom shifts between heteroatoms, Table 2 summarizes an analysis of the variation in predicted log K_{oc} and log S_w as compared to the number of tautomers per compound. For K_{oc}, 70% of the compounds (5198 compounds) yield variations across tautomers smaller than 0.5 log units, and 5% (378) yield log K_{oc} variations larger than 1. With water solubility, 80% (6237) of the compounds show variations across tautomers below one log unit, 6% log S_w variations between 1 and 2, and 1.7% log S_w variations larger than 2.

Interestingly, there is no distinct relationship between the number of tautomers per compound and the property variation across tautomers for both log K_{oc} or log S_w. Among the (many) compounds with less than five different tautomers, variations of more than two log units are still observed for 21 compounds with regard to K_{oc} and for 132 compounds with regard to S_w. Moreover, 33 compounds with less than five tautomers yield predicted S_w differences of more than four log units, 24 compounds with 6−50 different tautomers provide similar log S_w ranges, but none of the 28 compounds with more than 51 distinct tautomers yield log S_w variations above four log units.

Energetic Stability of Tautomers. Finally, we come back to the InChI approach of confining the H shifts to heteroa-

toms. As mentioned above, the reasoning behind this approach is the assumption that in most cases, the most stable prototropic tautomers are those generated through H shifts across heteroatoms.

For the nine compounds of Table 1, we have augmented these heteroatom-based tautomers manually by tautomers with H atom shifts involving sp^2 carbons. Moreover, AM1 has been applied to calculate their heats of formation, ΔH_f. As can be seen from comparing the third and the second last column of the table, inclusion of tautomer-active sp^2 carbons increases the total number of tautomers by 28−60. The associated increase in ΔH_f variation is between 0.3 and 82.1 kJ/mol, and above 10 kJ/mol in eight of the nine cases. Interestingly, for only two compounds (primisulfron and tribenuron), the lowest-energy tautomer is one involving an sp^2 carbon as H atom acceptor. For all other seven compounds, the subset of heteroatom-confined tautomers includes the thermodynamically most stable structure.

While the currently selected nine compounds are certainly not representative for the chemical domain of large inventories, such as the EINECS list, the results nevertheless suggest that in accord with the InChI approach, H atom shifts across heteroatoms will usually dominate the tautomeric equilibrium of chemical substances. Whether this can be confirmed with a large data set is subject of an ongoing study and will be reported in due course.

CONCLUSIONS

The presently developed algorithm enables a fast generation of all formally valid prototropic tautomers resulting from the mobile H atoms and their associated heteroatom acceptor sites as defined in the InChI code. As such, it provides an efficient tool to identify those tautomers that because of their restriction to H atom exchanges between heteroatoms usually include the thermodynamically most stable structures. Moreover, the energetic stability range of the accordingly generated tautomers is typically (but not necessarily) more narrow than when including sp^2 carbons as tautomer-active sites. Apart from energetic considerations, the algorithm yields a lower bound of the theoretically possible number of tautomers. If combined with more comprehensive tautomer generation procedures, the algorithm will allow for a fast assessment of the number of heteroatom-confined tautomers as compared to the total (or essentially total) number of possible tautomers. In addition, the developed tool may serve to screen the dependence of QSAR predictions, as well as of chemical database searches and associated structural retrievals on tautomeric features, and may provide pertinent information about respective variations across thermodynamically reasonable tautomers.

ACKNOWLEDGMENT

This work was partly funded by the EU projects OSIRIS (www.osiris-reach.eu, contract no. 037017), CAESAR (www.caesar-project.eu, contract no. 022674), and 2-FUN (www.2-fun.org, contract no. 036976), and by the German Federal Environmental Agency Umweltbundesamt (UBA). Moreover, the authors thank Barbara Wagner, Daniel Exner, Torsten Bloi, Dominik Wondrousch, and Daniel Stosch for their valuable help and technical support.

REFERENCES AND NOTES

(1) Harańczyk, M.; Gutowski, M. Quantum Mechanical Energy-Based Screening of Combinatorially Generated Library of Tautomers. TauTGen: A Tautomer Generator Program. *J. Chem. Inf. Model.* **2007**, *47*, 686–694.
(2) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W. D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2342–2354.
(3) Ghosh, A.; Wondimagegn, T.; Nilsen, H. J. Molecular Structures, Tautomerism, and Carbon Nucleophilicity of Free-Base Inverted Porphyrins and Carbaporphyrins: A Density Functional Theoretical Study. *J. Phys. Chem. B.* **1998**, *102*, 10459–10467.
(4) Zhang, Y. A.; Monga, V.; Orvig, C.; Wang, Y. A. Theoretical Studies of the Tautomers of Pyridinethiones. *J. Phys. Chem. A.* **2008**, *112*, 3231–3238.
(5) Kalliokoski, T.; Salo, H. S.; Lahtela-Kakkonen, M.; Poso, A. The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49*, 2742–2748.
(6) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Recept. Signal Transduction* **2003**, *23*, 361–371.
(7) Martin, Y. C. Let's Not Forget Tautomers. *J. Comput. -Aided Mol. Design* **2009**, *23*, 693–704.
(8) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M.; Delany, J. J. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161–1172.
(9) Trepalin, S. V.; Skorenko, A. V.; Balakin, K. V.; Nasonov, A. F.; Lang, S. A.; Ivashchenko, A. A.; Savchuk, N. P. Advanced Exact Structure Searching in Large Databases of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 852–860.
(10) Harańczyk, M.; Gutowski, M. TauTGen. http://tautgen.sourceforge.net (accessed Sep 15, 2008).
(11) Harańczyk, M.; Puzyn, T.; Sadowski, P. ConGENER—A Tool for Modeling of the Congeneric Sets of Environmental Pollutants. *QSAR Comb. Sci.* **2008**, *27*, 826–833.
(12) Milletti, F.; Storchi, L.; Sforna, G.; Cross, S.; Cruciani, G. Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases. *J. Chem. Inf. Model.* **2009**, *49*, 68–75.
(13) IUPAC (International Union of Pure and Applied Chemistry). IUPAC—International Chemical Identifier, InChI version 1.02 (beta). www.iupac.org/InChI/ (accessed Jun 3, 2008).
(14) Harańczyk, M. ConGENER. http://congener.sourceforge.net (accessed Sep 15, 2008).
(15) *ACD/ChemSketch*, version 11.01; Advanced Chemistry Development, Inc.: Toronto, ON, 2009.
(16) *MN.TAUTOMER*, version 1.8; Molecular Networks GmbH: Erlangen, Germany, 2007.
(17) Karickhoff, S. W.; Carreira, L. A.; Hilal, S. H. SPARC Performs Automated Reasoning in Chemistry. http://sparc.chem.uga.edu/sparc/ (accessed May 1, 2009). Developed at the University of Georgia through grants from U.S. Environmental Protection Agency.
(18) Molecular Networks GmbH. MN.TAUTOMER—Enumeration of Tautomers. http://www.molecular-networks.com/products/tautomer (accessed Sep 15, 2008).
(19) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introducing to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(20) CAS (Chemical Abstract Service). CAS REGISTRY and CAS Registry Numbers. http://www.cas.org/expertise/cascontent/registry/regsys.html (accessed Mar 1, 2010).
(21) Tavakol, H.; Arshadi, S. Theoretical Investigation of Tautomerism in N-Hydroxy Amidines. *J. Mol. Model.* **2009**, *15*, 807–816.
(22) Project: InChI and InChIKey: Further Promotion. http://www.iupac.org/web/ins/2008-033-1-800 (accessed Sep 25, 2009); continuation of projects 2004-039-1-800 and 2000-025-1-800.
(23) U.S. Secretary of Commerce. NIST Chemistry WebBook. http://webbook.nist.gov/ (accessed Mar 3, 2010).
(24) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
(25) Schüürmann, G.; Ebert, R.-U.; Nendza, M.; Dearden, J. C.; Paschke, A.; Kühne, R. Predicting Fate-Related Physicochemical Properties. In *Risk Assessment of Chemicals. An Introduction*, 2nd ed.; van Leeuwen, K., Vermeire, T., Eds.; Springer Science: Dordrecht, The Netherlands, 2007; pp 375−426.
(26) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(27) *MOPAC 93*, revision 2; Stewart Computational Chemistry: Colorado Springs, CO, 1994.

(28) Schüürmann, G.; Ebert, R.-U.; Kühne, R. Prediction of the Sorption of Organic Compounds into Soil Organic Matter From Molecular Structure. *Environ. Sci. Technol.* **2006**, *40*, 7005–7011.

(29) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Model Selection Based on Structural Similarity—Method Description and Application to Water Solubility Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 636–641.

(30) European Commission. ESIS (European Chemical Substances Information System). European Inventory of Existing Commercial Chemical Substances (EINECS). http://ecb.jrc.ec.europa.eu/esis/index.php?PGM=ein (accessed Sep 1, 2008).

(31) Daginnus, K. EC Chemical Inventories. http://ecb.jrc.ec.europa.eu/qsar/information-sources/ec_inventory (accessed Aug 5, 2008).

(32) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(33) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Chemical Domain of QSAR Models From Atom-Centered Fragments. *J. Chem. Inf. Model.* **2009**, *49*, 2660–2669.

CI1001179