J Comput Aided Mol Des (2010) 24:485–496
DOI 10.1007/s10822-010-9329-5

So you think you understand tautomerism?

Roger A. Sayle

Received: 1 February 2010 / Accepted: 10 March 2010 / Published online: 23 March 2010

Abstract It appears so simple at first glance, "tautomers are isomers of organic compounds that readily interconvert, usually by the migration of hydrogen from one atom to another". If a chemist can describe the problem so succinctly, one might question why the complication of tautomerism remains a considerable challenge to cheminformatics and computer-assisted drug design. With a half-century of experience with representing molecules in computers, and almost limitless modern computational power, the problem should have been solved by now. The unfortunate answer is that the frustration and inconvenience of a database search failing to find matches due to differences in the tautomeric forms of the query and registered compounds is but the tip of an iceberg. Prototropic tautomerism, the movement of hydrogens around a molecule, is but just one aspect of an interconnected web of complications. These include mesomerism, aromaticity, protonation state, stereochemistry, conformation, polymerization, photostability, hydrolysis, metabolism and EOCWR (explodes on contact with reality). The common theme is that valence theory, which underlies all modern chemical informatics systems, is an approximate theoretical model for representing molecules mathematically, and, as with all models, it has limitations and domains of applicability. In the physical environments that chemists care about, small organic molecules are often dynamic, existing in multiple equivalent or interconvertible forms. A single connection table can at best represent a snapshot or sample from these populations. Although partial algorithmic solutions exist for handling the most common cases of tautomerism, this perspective hopes to argue that the underlying problems perhaps make tautomerism more complex than it might first appear.

Keywords Tautomer · Tautomerism · Mesomerism · Protonation state · Enumeration · Resonance · Aromaticity

R. A. Sayle (✉)
NextMove Software, 1303 Bartlet Court, Santa Fe, NM 87501, USA
e-mail: roger@nextmovesoftware.com

Introduction

The field of molecular chemistry owes a debt of gratitude to the work of August Kekulé (1829–1896) and Gilbert N. Lewis (1875–1946) for establishing the structural formula as a mathematical model of chemical structure [1]. The representation of a chemical as an undirected graph, with atoms represented by vertices labeled by elements of a given valence, and bonds represented by edges annotated by bond order, revolutionized chemistry. The enormous influence of this view of chemistry cannot be overstated. Indeed, the connection tables and line notations of modern chemical information systems owe their origins to inventors who could not have foreseen the development of today's computers.

Particularly striking is that the Lewis/Kekulé structure is the foundation of almost all modern chemical informatics systems, even though more advanced and more accurate theoretical models of chemistry, such as quantum mechanics, have been available for some time. The clue to the longevity of valence theory is its computational tractability. Although quantum mechanical representations, such as molecular orbital theory and valence bond theory, are universally acknowledged to be more faithful models of chemistry, they are considerably harder models to encode and reason about computationally. Perhaps the single greatest attribute of valence theory is concept of "identity",

followed closely by the less well-defined concepts of substructure, superstructure and similarity. Given two graphs representing molecules, it is a well-defined (and well studied) task to ask whether the two graphs are the same (termed isomorphic in graph theory). Alas, quantum mechanics does not have a comparable notion of equality; one can confirm that two "systems" have the same number of nuclei, with equivalent numbers of protons and neutrons, and that they have the same number of electrons with the same total spin, but each 3D configuration of nuclei gives rise to a different set of wave functions. In most quantum mechanical formulations, there's no distinction between conformers, tautomers and structural isomers. Every arrangement in space is no more or less unique than any other. The development of next generation cheminformatics systems based on quantum mechanics rather than valence theory would mark a major advance for the field.

Classic tautomerism

When many chemists and authors discuss tautomerism, they often restrict their consideration to simple cases of prototropic tautomerism, giving frequent examples based on nitrogen-containing aromatic heterocycles. These systems, such as the annular tautomerism shown in Fig. 1, are useful for introducing students to the fundamental issues [2–4]. Under physiological conditions, a compound such as 4-pyrimidone exists as a population of rapidly interconverting tautomers. Figure 1 shows the two major species.

However, in addition to the two annular forms shown in Fig. 1, there also exist additional minor species. The next most obvious is the enolic form, caused by prototropic migration of the pyrrolic hydrogen to the ketone oxygen, as shown in Fig. 2.

In Fig. 2, I've intentionally drawn both Kekulé forms of pyrimidin-4-ol to highlight the relationship between tautomerism and resonance. The aromatic nature of pyrimidine means that there exist two distinct Lewis structures for 4-hydroxypyrimidine, such that a pure valence theoretical representation is not unique or canonical. This is a well known and understood complication in cheminformatics that is readily solved by the introduction of aromatic bond types or canonical Kekulé forms. But the principal is the

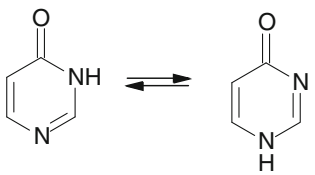


Fig. 1 Annular tautomerism of 1H- and 3H- forms of 4-pyrimidone

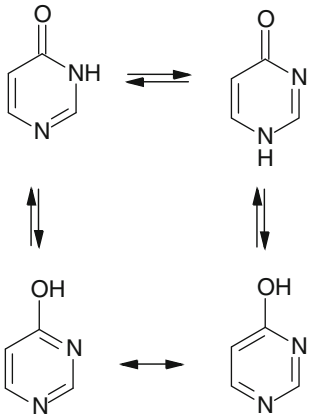


Fig. 2 The three major tautomeric species of pyrimidin-4-one. The bottom two structures are different Kekulé representations of the same enolic form, indicated by the *single double-ended arrow* rather than the *pair of arrows* denoting an equilibrium

same for tautomerism; where all of the Lewis structures in Fig. 2 denote a single chemical entity such as a compound dissolved in water or a drug circulating inside a patient. A classic historical perspective on this "two sides of the same coin" is given by Linus Pauling, in the section "The relation between resonance and tautomerism" in his book "The Nature of the Chemical Bond" [5] (appropriately, as mentioned as the start of this article, dedicated to Gilbert Newton Lewis).

This example also highlights a second important point, that the simplification of a system to its major species is an approximation. The two major tautomeric forms considered in Fig. 1, are a reduction of the three tautomeric forms (four resonance forms) in Fig. 2, which in turn are a reduction of a much larger number of significant forms or states available to 4-pyrimidinone in dilute aqueous solution. It is unfortunately not uncommon for schemas given in papers on tautomerism to fail to mention, let alone identify, the predominant species for the conditions under discussion.

Tautomerism, mesomerism and ionization

In addition to prototropic tautomerism and aromaticity, there are numerous additional complications when attempting to use Lewis structures to judge the molecular identity of a compound under physical conditions [6]. The most significant of these are ionization or disassociation (the loss or addition of protons in aqueous solution due to the acid and base natures of a compound) and mesomerism (the need to place formal charges on particular atoms to represent charged systems). The interconnected nature of tautomerism, ionization and mesomerism can be shown via the schema shown in Fig. 3.

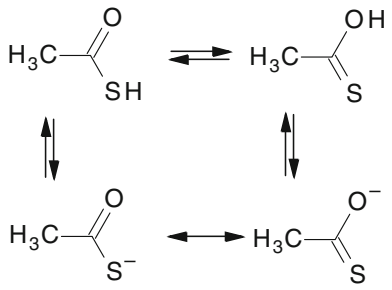


Fig. 3 The relationship between tautomerism (*left–right at top*), ionization (*top-bottom*) and mesomerism (*left–right at bottom*) from the equivalent forms of thioacetic acid

All four Lewis structures in Fig. 3 denote the same compound, thioacetic acid or thioacetate, and would be expected to be represented by a single entry in a chemical registration system, for example identifier 10-08 in Chemical Abstract's registry. The top two neutral forms are clearly tautomers of each other. However, the pKa of thioacetic acid is about 3.33 (at 25 °C) so in aqueous solution at room temperature it would predominantly exist as the bottom form, which are resonance forms of each other. Unlike carboxylates with two symmetric oxygen atoms, thiocarboxylates have two distinct Lewis structures, even though quantum mechanically (physically) the electrons are not associated with one particular atom or another.

Much like the use of aromatic bonds to avoid duplicate Kekulé forms, most chemical database systems also encode some level of normalization or "business rules" to prevent ionization duplicates (both carboxylic acids and carboxylates) and mesomeric duplicates (such as alternate forms of functional groups, such as nitros and azides as shown in Fig. 4).

Typically, software for registration in a database system will attempt to neutralize a molecule, stripping protons from charged amines and adding protons to "olates" in order to give a molecule no net charge, preferably without zwitterions. For virtual screening, however, software may prefer to normalize molecules to a plausible ionization state around pH 7. Correspondingly, the preferred valence representations of functional groups (such as hypervalent or charge separated forms of nitro groups) are identified by pattern matching alternate forms, and replacing them with their canonical preferred representation. In both processes, there is often no right answer. Mesomeric forms are artifacts of valence theory, so choice of "preferred" styles are purely a matters of convention or style, with different pharmaceutical companies or vendor catalogues often adopting different (competing) aesthetics. Likewise, for protonation state and tautomeric forms, compounds frequently exist in two (or more) forms, with neither being significantly more popular than another. Provided that the mechanism used to select the representative form is consistent, duplicates can be easily identified.

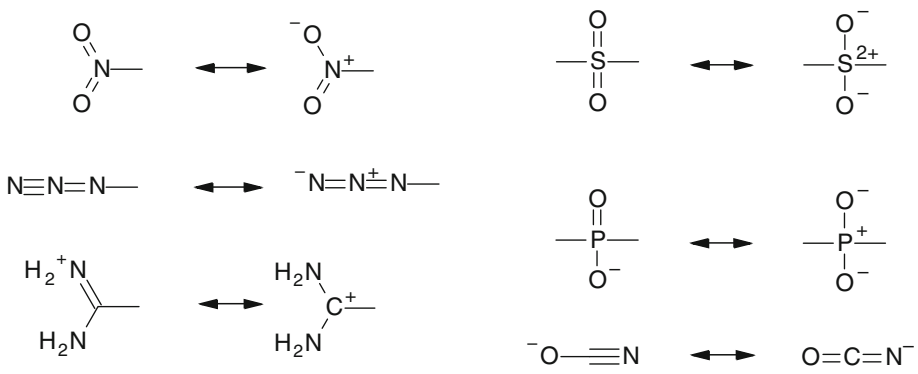
The real complexity starts to occur when each of these mechanisms start overlapping, and the distinctions between them start to blur. Figure 5 shows four "equivalent" Lewis structures that present a difficult challenge for normalization software.

When presented with either neutral form of a histidine sidechain, an active site preparation program might to decide to protonate the imidazole ring system reflecting physiological pH. However, a choice then has to be made as to where to locate the resulting formal charge, as shown in Fig. 6. Although a protonated histidine (HIP) residue isn't a tautomeric problem, the two distinct Lewis structures can potentially create problems.

Similarly, when registering a substituted guanidinium into a chemical database, a program may decide to neutralize the molecule, but by breaking the symmetry now has to decide where to place a double bond (as shown in Fig. 7).

An excellent real-world demonstration of the difficulty in recognizing "equivalent" forms of the same compound is given by Paul Labute [7] as shown in Fig. 8. In his comparison of algorithms for perceiving molecules from 3D atomic co-ordinates, he depicts the two structures of methazolamide, recognized from the co-ordinates of MZM in the PDB file 1bzm, arguing that the one on the left is

Fig. 4 Functional group mesomers normalized during compound registration



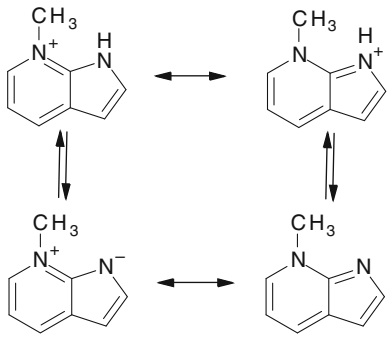


Fig. 5 Four equivalent (mesomerism and ionization) forms of the same compound

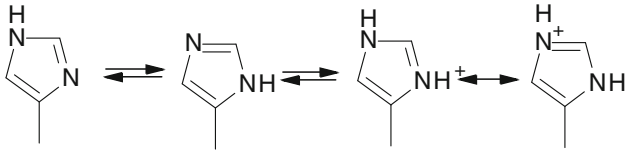


Fig. 6 Neutral and protonated forms of a histidine sidechain (imidazole/imidazolium)

correct and the one on the right is wrong. As should be clear from the previous discussions, these are both equivalent ionization states of the exact same molecule, with neither being inherently more correct than the other. It should be noted that in the given reference, Labute depicts only a single nitrogen in the thiadiazole ring which is inherently incorrect, as this has the wrong molecular formula. If anything, methazolamide has an experimentally measured pKa of 7.3, making the protonated form (on the right) more common at physiological pHs.

Another thought provoking example of blurred distinction between tautomerism and mesomerism is given by the unusual representation of sulfonic acid shown in Fig. 9. An analysis of the National Cancer Institute's August 2000 compound database looking for functional groups that could not be named by OpenEye's Lexichem [8] software discovered this SO_3H representation as the most frequent unusual functional group. Upon first inspection, this appears to be a tautomer issue with a proton moved from one oxygen to another, but on closer analysis this is actually a mesomer issue. The formal charges and bond orders are artifacts of valence theory. Hence the arrow in Fig. 9 is a single double-headed arrow rather than a pair of arrows, representing an equilibrium in this work.

Another interesting example from the NCI database is the representation used for pyridine *N*-oxide, shown in Fig. 10. This is a case of simple mesomerism, but it demonstrates that representations that might be considered highly unusual by many chemists do occur in real-world databases (surprisingly frequently). With the huge number of chemical structures in modern databases, it is impractical to check and verify each by hand. In many cases, problematic structures have been generated by combinatorial library software or similar automated processing, resulting in chemistries that would be unlikely to be drawn (that way) by hand.

Interesting cases

An interesting form of tautomerism is shown if Fig. 11 where the migration of the proton results in a change of atomic valence of a non-terminal atom. In most examples of tautomerism, the sum of bond orders (including those to hydrogens) is invariant under the transformation/equivalence.

Another further example of this valence changing form of tautomerism is exhibited by interconversion of phosphinic acids and hypophosphorous acids, as shown in Fig. 12. Perhaps counter-intuitively, the major tautomeric form of these compounds is the hypervalent form.

A convenient property of most examples of tautomerism is that atomic hybridization of each non-terminal atom is preserved. This conservation of local geometry at each

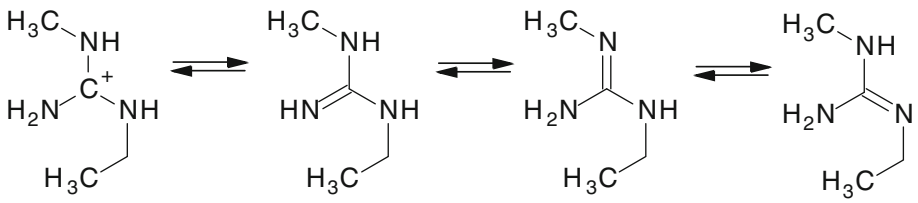


Fig. 7 Neutralizing guanidinium (*left*) can lead to three different tautomeric forms (*right*)

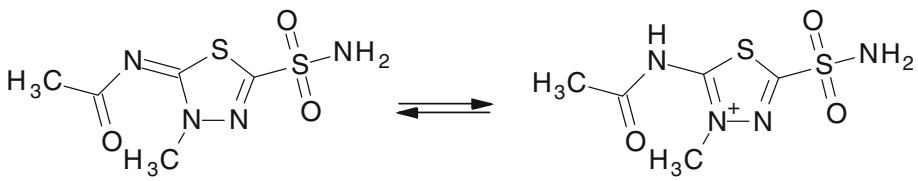


Fig. 8 Equivalent forms of methazomalide

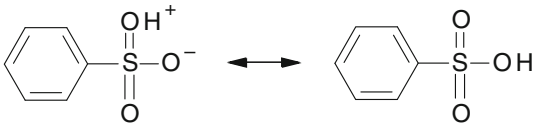


Fig. 9 Strange crypto-tautomeric representation of benzenesulfonic acid found in the NCI database

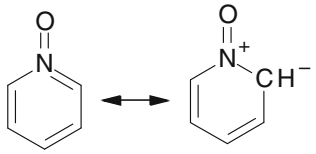


Fig. 10 Representation of pyridine *N*-oxide in the NCI database

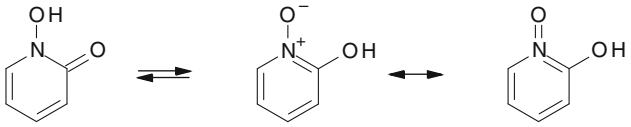


Fig. 11 Example change of non-terminal atomic valence between prototropic tautomers

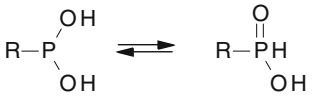


Fig. 12 Valence changing tautomerism of hypophosphorous acids

atom (trigonal planar for sp^2, tetrahedral for sp^3, etc.) means that tautomers forms of molecules have highly similar conformations, and low RMS values for heavy atom superpositions. This general property provides an algorithmic advantage of the 3D analysis of tautomers, where the co-ordinates of one can be used as an approximation or starting point for the co-ordinates of another. The application of this technique in structure-based drug design (docking) is discussed by Sayle and Nicholls [9]. The principle is that a single representative tautomer may be posed in a protein active site, and its family of tautomers rapidly scored using the same pose, rather than the more conventional approach of enumerating all tautomeric forms first, and then performing conformation generation and docking on each independently.

However, although the conservation of geometry property applies to most forms of tautomerism, it remains a heuristic rule-of-thumb with several classes of counter examples. These include "C-type" systems (where the proton moves from or to a carbon atom [10]), as shown in Fig. 13, and quaternary amine systems (where the movement of the hydrogen to a secondary or tertiary amine induces tetrahedral geometry).

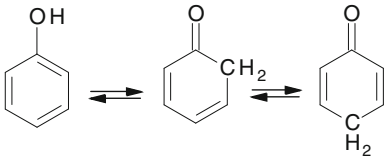


Fig. 13 Tautomeric forms of phenol, showing c-type tautomerism

The interaction between tautomerism, mesomerism and ionization with c-type tautomers can also be helpful and educational in rationalizing "acidic carbon" atoms. In some classes of organic acid/base systems, such as tetronic acids and barbiturates, some medicinal chemists may be misled into believing the experimental pKa of a molecule is the result of a loss a proton from a methylene carbon atom. A clearer picture is revealed in the Fig. 14, which explains that the proton is actually lost from an oxygen atom in one of the tautomeric forms. Tetronic acid (tetrahydrofuran-2,4-dione) has an experimental pKa of about 3.7, and is therefore anionic at physiological pHs [11, 12].

In addition to the C-type tautomerism explained above, in which a carbon interconverts between sp^2 and sp^3, there also exist examples of tautomerism through sp-hybridized (Fig. 15) and sp^3-hybridized centers (Fig. 16).

The sp^2 example in Fig. 16, also demonstrates the potential interaction between tautomerism (or mesomerism) and stereochemistry [13, 14]. Based upon the Lewis structure, the central phosphorus atom may potentially be incorrectly perceived as a chiral center, even when X or Y are hydroxyl, hydroxylate, thiol or thiolate groups. Likewise, tautomers that conjugate through acyclic double bonds may have problems preserving/annotating cis vs. trans configurations without the ability to represent IUPAC's s-*cis* and s-*trans* forms of stereochemistry (such as of buta-1,3-diene) [15].

Although sp^3-hybridized hetero-center tautomers may seem artificial, they really do occur in practice as registration complications in large chemical databases. The

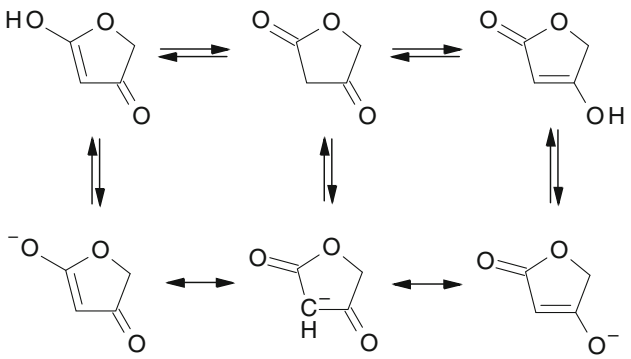


Fig. 14 The mythical acidic carbon atom of tetronic acid

Fig. 15 Tautomerism through a sp-hybridized center

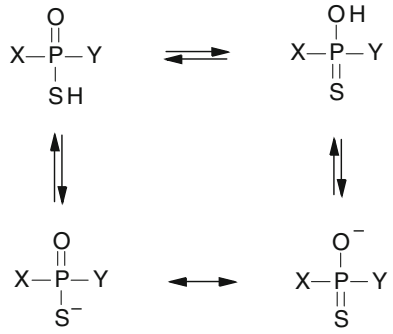


Fig. 16 Tautomerism through a sp^3-hybridizied center

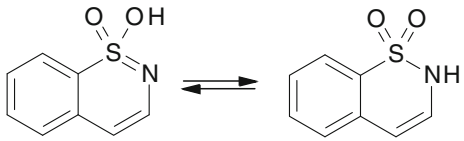


Fig. 17 A "real-world" example of a sp^3-hybridized center tautomeric duplicate

compound in Fig. 17 was discovered by searching a large on-line chemical database for tautomeric duplicates.

Local versus global approaches

Algorithmic solutions to handling tautomerism may be loosely categorized into two broad classes; local methods and global methods. The local class of techniques is based upon pattern matching, encoding rules that apply to small substructures. These methods are easy to implement and catch the majority of important cases. Examples of this sort of approach include the Intervet CACTVS rule set [16], which contains 21 patterns, AstraZeneca's Leatherface rule set [17] which contains 140, and the TauThor system of Milletti et al. [18] which iterates a single rule repeatedly, and the database search technique of Trepalin et al. [19] which only handles 1,3-tautomerism and annular tautomerism in 5-membered rings and aromatic carbocycles. Additional examples of local methods can be found in the literature [20, 21].

One limitation with local tautomerism rules is that they fail to identify the equivalence of long-range tautomer pairs, where the proton migrates a significant distance. A common case of this are tautomers formed by zwitterions. Figure 18 shows prototropic isomerism in a dialanine

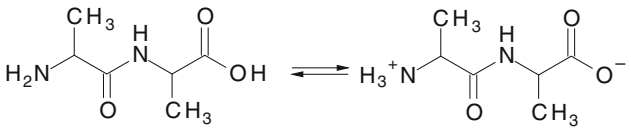


Fig. 18 Prototropic (zwiterionic) tautomerism of dialanine (DL-alanylalanine)

dipeptide. Clearly, the two forms are tautomeric forms of each other, differing only in the migration of a single proton, and is a model for much larger peptides and proteins. This example demonstrates that the migrating proton may move between atoms separated by large distances, but topologically and geometrically, and not even be part of the same conjugated system.

Another class of "long range" tautomerism involves the equivalence of salt forms. A complicating factor when maintaining a database of the marketed/prescribed forms of drugs is the handling of the salt forms of each compound. As implied by the term salt, the formulation of a therapeutic compound often consists of acidic and basic components, whereby the movement of a proton between them creates the salt. Curation of these databases requires care in how these salt forms are registered, especially if attempts are made to prefer physiological representations. Figure 19 shows two equivalent forms of the antiplatelet drug ticlid.

Global tautomerism approaches tackle the problem in a more holistic manner. In much the same way that computational quantum chemistry codes determine the locations of electrons around a configuration of nuclei, global tautomer algorithms place a specified number of protons on a topological scaffold of heavy atoms. And in much the same way as electrons populate the more energetically favorable orbitals first, the protons associate with the most favorable heavy atoms. One of the simplifying principles of quantum chemistry is the Born–Oppenheimer approximation, that states that electrons reorganize around atomic nuclei fast enough to consider the nucleic fixed and the electronic reorganization instantaneous. The author proposes a variation of this principle for tautomerism, that protons and electrons reorganize around heavy atom nucleic fast enough to consider the heavy atoms fixed and the proton/electronic reorganization instantaneous.

Global approaches to tautomerism necessarily identify a superset of tautomers to those found by local approaches, potentially allowing hydrogens and formal charges to

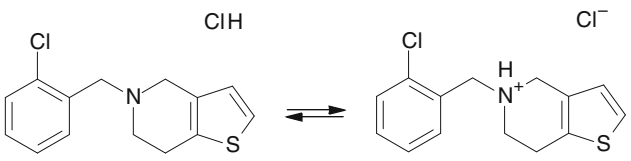


Fig. 19 Tautomeric salts of ticlid (ticlopidine hydrochloride)

migrate large distances. Example implementations of global tautomerism algorithms include OpenEye Scientific Software's tautomers [22, 23] and Accelrys' Pipeline Pilot enumerate tautomers component [24].

Comparing the results of "systematic" global algorithms to "pattern matching" local algorithms reveals a number of strengths and weaknesses to each approach. The use of specific (often hand-coded) patterns and rules in local methods produces few surprises, the tautomers and mesomers that are found are those that are looked for. Global techniques, on the other hand, are capable of identifying many obscure and delightful equivalences, never considered by most pattern libraries. The surprises whilst a benefit when searching for compounds, are often less appreciated during enumeration or canonicalization.

In an analysis of a corporate registry system, global equivalence testing was able to identify the class of mesomeric equivalence shown in Fig. 20 that had not been caught by the extensive list of in-house business rules. Once identified, it is possible to add an additional pattern-based local rule to prevent such duplication in future.

The five computations

Whether we are talking about tautomerism, mesomerism or ionization state, a researcher may be interested in computationally solving one of five types of problem. These five tasks are common to all fields where one is dealing with populations of representations that occur with different frequencies/energies.

1. *Comparison*. Given two molecules can we determine that one is a tautomer of the other?
2. *Canonicalization*. Given a molecule can we generate a unique encoding of its set of tautomers, such that the encodings of two molecules are identical if and only if those two molecules are tautomers of each other?
3. *Enumeration*. Given a molecule can we list all of the molecules that are tautomers of it?
4. *Selection*. Given a molecule can we list a subset of its energetically most likely tautomers, given a particular environment (solvent, temperature, pH, binding site, etc.)?
5. *Prediction*. Given a molecule can we predict its energetically most likely set of tautomers with their ratios, given a particular environment (solvent, temperature, pH, binding site)?

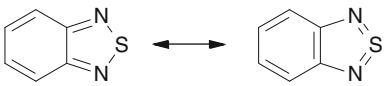


Fig. 20 Mesomeric duplicates of benzothiadiazoles

The first three are conceptually discrete cheminformatics problems that may have a well defined solution, but may differ based upon the operational definition of tautomer. The last two are computational chemistry problems that require some form of floating point energy calculation. The first three are problems in computer science; the last two are problems in chemical physics and physical chemistry.

Canonicalization (#2) is technically a superset of comparison (#1); a solution to either is sufficient for compound registration or duplicate removal, though canonicalization allows for more efficient implementation. An intermediate between these two is the use of tautomeric hash codes, that can be used to speed up comparison, but that do not have to be unique. With canonicalization, there is also a choice as to whether the canonical representation is a valid molecule, such as a representative member of the set of tautomers, or a more abstract encoding, such as chosen by IUPAC's InChI identifier encoding. Of course, schemes that select a representative canonical form do not need to select a physically reasonable tautomer. Any tautomer will do provided it is selected consistently, indeed some methods choose the alphabetically first SMILES when sorted lexically [17].

Full enumeration (#3), although the most common application of tautomer software, is probably the most problematic. The more inclusive the operational definition of tautomerism is, the much larger the number of potential tautomers. Many molecules of pharmaceutical (or dye) industry interest have many thousands to many millions of tautomeric forms. The goals chemical registration (tasks #1 and #2) are fundamentally opposed to those of enumeration. A capable registration system should be able to spot that both representations in Fig. 21 denote the same molecule, but it might be unreasonable for an enumeration program to return the first as a potential form of the second.

Ultimately, enumerating all possible tautomers is a futile task, impractical for even moderately sized molecules. As a result, evaluating or comparing tautomer software purely by the number of exhaustively enumerated tautomers they can generate is of limited value. Instead, the intelligence of tautomerism software lies not only in the tautomeric forms it generates, but also in the tautomeric forms it discards (task #4).

Selection (#4) and prediction (#5) distinguish themselves from the earlier (perhaps simpler) tasks by relying on a scoring scheme for ranking tautomeric forms by

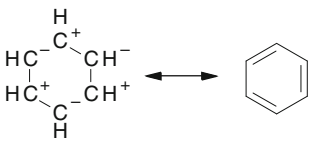


Fig. 21 Equivalent mesomeric forms of benzene, highlighting conflicting aims of registration and enumeration

likelihood. These two tasks are distinguished by the quality of the score evaluation. For some applications, a simple triage of plausible versus implausible may be sufficient. For others, a holy grail of the field would be to quantitatively estimate the expected ratios or energy differences between tautomers.

Unfortunately, the very small energy differences between tautomeric forms make them very difficult to accurately calculate. Methods such as quantum mechanics, or atom type based heat of formation/heat of solvation calculations involve the subtraction of two large numbers, to produce a result typically smaller than the expected computational error. Moran et al. [25] presents a state-of-the-art high-level quantum mechanical calculation of the 2-pyridinethiol/thione system describing the significant computational effort required to reproduce the results encoded by even simple rule-based systems.

Worse still the small energy differences that give rise to tautomeric preference can easily be dwarfed by environmental influences, such as choice of solvent or interaction in a protein active site. The influences of local charged groups in a protein active site can completely overwhelm any subtle preferences a molecule may have in solvent or vacuum. Flipping a tautomeric form to produce two complementary hydrogen bonds, where previously there was both a donor-donor clash and an acceptor-acceptor clash, is energetically so favorable that the induced tautomer might never be observed in bulk solvent. The literature is replete with examples; methotrexate binds to dihydrofolate reductase (DHFR) as the protonated (at N1) form, even though it prefers to be neutral in water. The active site histidine residues in zinc binding proteins are typically negatively charged imidazole anions, even though the pKa for that proton loss is about 14.5, or over 7 log units from physiological pH. An excellent example is given by Kenny and Sadowski [17] who describe the difficulty in reproducing the conformation of indoline in the binding site of cyctochrome C peroxidase (PDB 1aek). Although their Leatherface software correctly determines that indoline is expected to be neutral in bulk solvent (pKa of about 4.9), the proximity to ASP235 is sufficient to induce/recognize the protonated form. Indeed, the title of the paper describing the X-ray crystal structure in pdb1aek, "characterization of an engineered heterocyclic cation-binding site", would appear to support the fact that indoline is bound as a "heterocyclic cation". Finally, an abundant source of examples of how environment can influence the preferred tautomeric form is the Cambridge Structural Database [26]. A study of the polymorphs and crystal packing in such small molecule crystallography databases reveals a significant number of instances where a tautomeric compound is observed in different tautomeric forms within the asymmetric unit of a unit cell. In these cases, the energetic penalty for adopting an alternate tautomer is overcome by the improvement in lattice energy for the resulting crystal packing.

Tautomeric preference and aromaticity

The previous sections have been careful to avoid (or limit) mentioning the principles by which one tautomeric form is preferred over another. Whilst this perspective article strongly argues for the need to develop tautomer "force-fields" or improvements in the methods of calculating Kt, the many terms and forces responsible for the phenomenon of tautomerism are beyond the scope of a single paper. Many greater minds than mine have struggled with Schrodinger's equation to figure out how to place electrons on a structure, developing breakthroughs like Hartree–Fock self consistent field (HFSCF) theory and density functional theory (DFT). I would not expect the harder task of placing both electrons *and* protons on a structure to be any easier. However, I'll take this opportunity to caution against "quick-fix" solutions, by drawing attention to a common misunderstanding (or incorrect assumption) in the field of tautomeric preference.

Researchers are now beginning to tackle the complex task of scoring tautomeric forms [27]. Alas a recurring feature of the functional forms being proposed, including those of Oellien et al. [16] and Milleti et al. [18] is a strong preference towards the aromatic forms of tautomers. Unlike double bond conjugation, aromaticity has relatively little influence upon the tautomeric preference of small organic molecules. In fact, many of the known examples of aromatic versus aliphatic annular tautomerism are driven more by bond energies, hydrogen bonding and geometrical constraints than by aromatic resonance energy.

The classic examples of tautomeric preference running counter to aromaticity are the hydroxyfurans (furanols), as shown in Fig. 22, though exactly the same effect is seen with thiophenes and pyrroles [28].

Other simple examples include indoles and isoindoles as shown in Fig. 23.

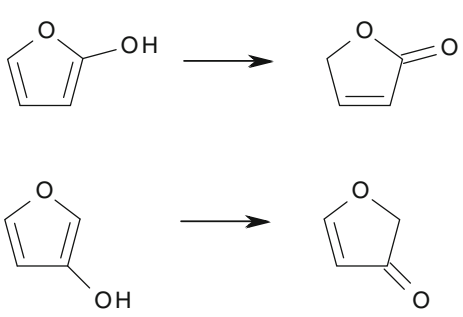


Fig. 22 Aqueous tautomeric preference on hydroxyfurans

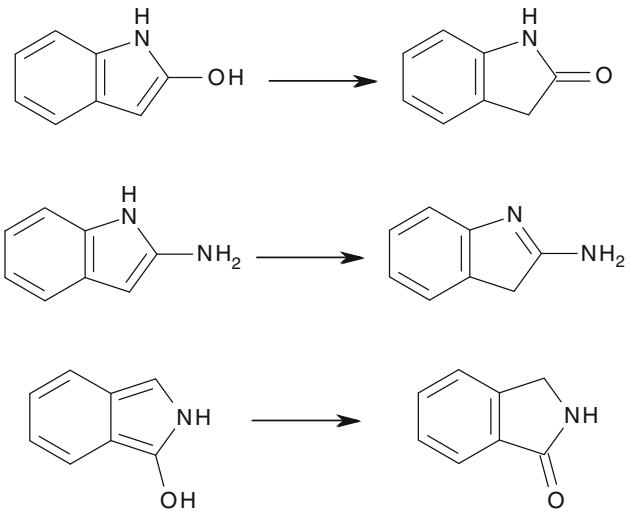


Fig. 23 Aqueous tautomeric preference of indoles and isoindoles

Similar things happen with 3-pyrazolones, where the 1,2-dihydro is less favored than the 2,4-dihydro form, as in Fig. 24.

In polar solvents, the rearrangement shown in Fig. 25 takes place. The form on the left is aromatic whilst the form on the right is not. In this case reversing the sense of the intra-molecular hydrogen bond from OH..N to NH..O is sufficiently strong to overcome any benefit of aromaticity, though notice that the atomic hydridization and therefore the geometry is preserved.

Another example that disrupts aromaticity involves cationic nitrogen heterocycles (Fig. 26).

The behavior of oxazoles and thiazoles differs in their preference to maintain aromaticity. Consider the two ring systems below where X is either oxygen (an oxazole) or sulfur (a thiazole) and R_1 and R_2 are π-acceptors (Fig. 27; such as COR, CO_2R, CN, NO_2, SO_2R, etc.…).

The tautomeric form on the left is aromatic, by many aromaticity models including those of OpenEye and Daylight, whilst the form on the right is not. It turns out that whilst the aromatic form is preferred for oxazoles (X=O), the non-aromatic form is preferred for thiazoles (X=S).

Another example is the case in Fig. 28 where several models of aromaticity consider the first form to be more aromatic than the second.

Finally, my last aromatic-related example is shown in Fig. 29. Although in the OpenEye/Daylight pi-electron counting scheme the left-hand rings on both sides have

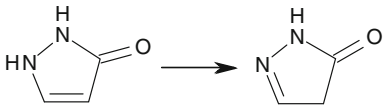


Fig. 24 Aqueous tautomeric preference of 4-pyrazolone

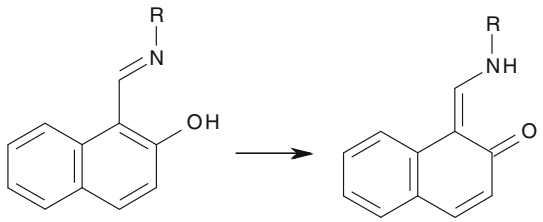


Fig. 25 Aqueous tautomeric preference for hydrogen bonding over aromaticity

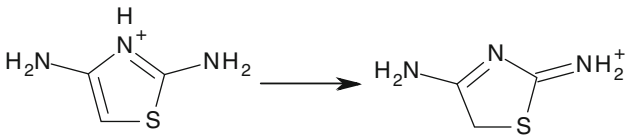


Fig. 26 Aqueous tautomeric preference for heteroconjugation over aromaticity

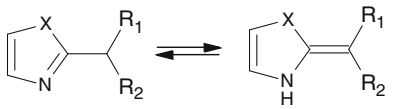


Fig. 27 Differences in aqueous tautomeric preference

only 4 electrons, and therefore are not aromatic, one might have thought that the amide form on the left to be better stabilized than the form on the right.

To summarize, a review by the author of published experimentally measured tautomeric ratios and tautomeric preferences involving both aromatic and non-aromatic forms reveals the vast majority of these equlibria prefer the non-aromatic form. This is true, for example, for the tautomer examples given in Katritzky's "Handbook of Heterocyclic Chemistry" [28]. This observation runs counter to all of the tautomer scoring functions described to date, where biasing for aromaticity potentially breaks more molecules than it fixes. Frustratingly, there remain counter-examples, such as the phenols in Fig. 13, where aromaticity is preferred, so one cannot just always blindly prefer the non-aromatic form instead.

It is the author's belief that in many of these examples, it is geometry and not aromaticity nor conjugation that accounts for the observed preference. In a six-membered ring, a carbon would prefer to adopt a trigonal planar sp^2 configuration with an internal bond angle of 120°. In a five-membered ring, a carbon atom prefers to adopt a tetrahedral sp^3 configuration with a strained internal bond angle of 109.5°. The hypothesis is that the energetic "bond-angle" strain dominates aromatic resonance energy. Extending this hypothesis, a pyrrole-like nitrogen (with a hydrogen) prefers five-membered rings, whilst pyridine-like and pyridinium-like nitrogrens prefer six-membered rings. In Tripos' Sybyl mol2 files these are termed N.pl3 and N.2 or N.ar

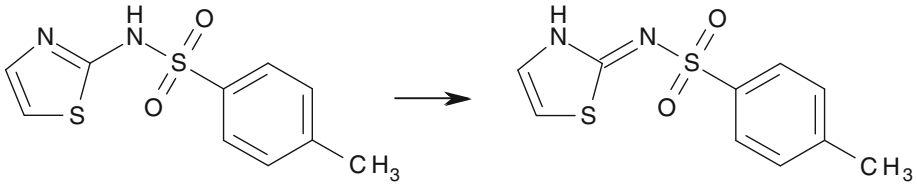


Fig. 28 Aqueous tautomeric preference for the less aromatic form

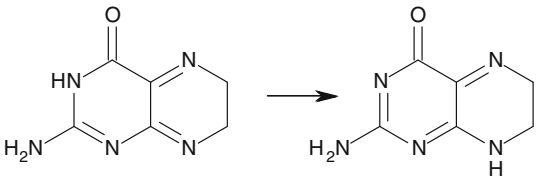


Fig. 29 A strange aqueous tautomeric preference

respectively. Such an interpretation predicts the tautomeric preference shown in Fig. 30.

The limited understanding of tautomeric principles by the general population of chemists represents perhaps one of the greatest dangers to the field. Software for handling tautomers may be dismissed by its users unless it returns the type of results that they expect to see. The reason why some tautomer enumeration programs prefer to generate aromatic tautomers, contrary to experimental evidence, is to satisfy customer demand and market forces, and not necessarily to produce the physically observed result.

Transition barriers

Once one starts down the slippery slopes of considering energetics in tautomerism, things slowly become even more complicated. In addition, to the ground-state energy of each tautomeric form is becomes necessary to consider the energy barriers between them. Given this barrier energy between tautomeric states and the physical conditions (including solvent, temperature, pressure, etc.), one can estimate whether the rate of conversion between tautomers is sufficient to consider them equivalent.

The significance of this issue can be understood by giving some thought to the commonly accepted definition of a tautomer. Two molecules are considered tautomers of each other if they interconvert by the movement of hydrogen atoms. Technically, under this definition, but-2-yne and buta-1,3-diene are tautomers of each other, as shown in Fig. 31. And if that seems bizarre, Fig. 32 shows that 1-(3-bromocyclohexyl)-3-chlorobenzene is technically a tautomer of 1-bromo-3-(3-chlorocyclohexyl)benzene. Whilst it these forms might conceivably equilibrate over geological or cosmological time scales, to medicinal and synthetic chemists they are distinct and separable.

By reductio ad absurdum, there must be something suspicious in our definition of tautomerism. Even if we restrict the above definition to "… the movement of a hydrogen atom", a little thought shows that both of the above pairs can be converted by a sequence of single proton relocations. The critical missing ingredient with the more pedantic definitions of tautomerism is the qualification of "interconvert". Two molecules are considered tautomers of each other if they readily interconvert by the movement of hydrogen atoms. Of course, the term "readily" then needs its own precise definition.

A proposal for a suitable definition has been suggested by Valters and Flitsch [29]. Consider two tautomers A and B of the same molecule, in Fig. 33. The tautomeric ratio, K_t, of their equilibrium is defined by the free energy difference between them, ΔG°. However the rate of equilibration is governed by the (lowest) free energy of activation between them, $\Delta G^{\neq}$. If ΔG° is greater than about 8–9 kcal/mole, the minor tautomer is unlikely to be observed, being less than about a millionth of the population. Valters and Flitsch propose a value of about 25 kcal/mole (corresponding to a rate of about 10^{-5} s^{-1}) as an arbitrary delineation (under normal conditions) to distinguish between a dynamic, rapidly attainable tautomeric equilibrium and a slower isomerization.

This view of tautomeric energy landscapes highlights both the potential asymmetry and non-reflexive nature of tautomer enumeration. Potentially, the enumerated set of tautomers of B may contain A, but the enumerated set of tautomers of A might not contain B. Likewise, the set of potential tautomers for a compound might not contain itself.

Fig. 30 Prefer N.pl3 in 5-membered rings and N.2/N.ar in 6-membered rings

Fig. 31 2-Butyne and 1,3-butadiene are tautomers of each other

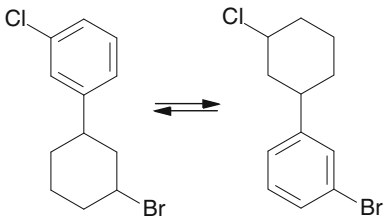


Fig. 32 Not all tautomers are readily interconvertible

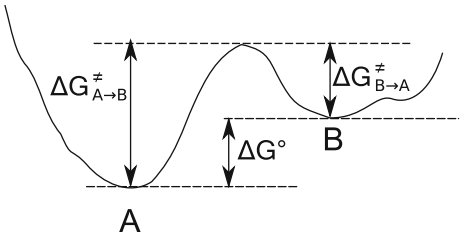


Fig. 33 Tautomer energy landscape of the equilibrium A↔B

Future work

By now the reader will forgiven for believing that the simple requests of "what forms does this compound have in a test-tube" and "what form can/will it adopt in an active site" are daunting. Unfortunately, the issues of tautomerism, ionization/disassociation, mesomerism and aromatic resonance are but the first in a long list of further complications. It is possible to draw impossible or highly strained molecules in a computer that just can't be made, and currently there are almost no tools that can distinguish these "explode on contact with reality" connection tables from true virtual compounds. Many compounds are reactive or unstable under typical conditions. Exposure to sunlight or water can result in unanticipated rearrangements. Figure 34 below shows the auto-polymerization of cyanic acid into cyanuric acid and isocyanuric acid which are tautomers of each other.

This perspective has avoided the discussion of ring chain tautomerism, such as the interconversion between chain and ring forms of glucose. But ring chain tautomerism is itself an example of a much larger field of reversible intermolecular reactions, which include transformations such as the Claisen rearrangement shown in Fig. 35.

Inside the body or living cells, things are even more complex as various enzymes transform xenobiotic

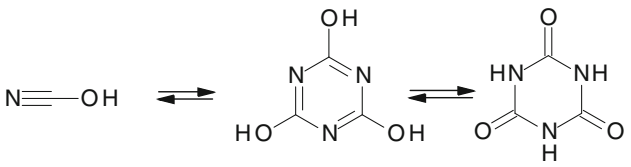


Fig. 34 The polymerization of cyanic acid into cyanuric acid and isocyanuric acid

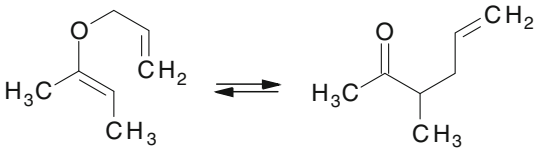


Fig. 35 An example of the Claisen rearrangement

molecules into various metabolites [11]. The entire field of prodrugs and active metabolites concerns the preparation of compounds that become equivalent at their point of therapeutic interaction.

The ray of hope amongst the dark litany of problems and pitfalls is that the pharmaceutical industry hasn't done too badly at improving human healthcare to date, without fully capturing or rationalizing all of the subtlety of the underlying chemistry. Perhaps mastery of tautomerism and resonance forms is a mostly academic challenge and not a critical path issue for rational drug design. This may be much like the field of aerodynamics, where the lack of understanding of how bees fly (until relatively recently [30]), has not prevented the aerospace industry, and companies like Boeing, from making many billions of dollars and putting a man on the moon.

Acknowledgments The author would like to acknowledge the patient mentoring on the complex subject of tautomers from Peter Taylor, Peter Kenny and John Bradshaw. I'd also like to thank Evan Bolton, Andrew Grant, Ben Ellingson, Jack Delany, Geoff Skillman, Jens Sadowski, Hugo Kubinyi and Yvonne Martin for many interesting and enlightening discussions and numerous perplexing tautomeric examples.

References

1. Lewis GN (1916) J Am Chem Soc 38(4):761
2. Baker JW (1934) Tautomerism. Van Nostrand, New York
3. Elguero J, Marzin C, Katritzky AR, Linda P (1976) Advances in heterocyclic chemistry. Supplement 1: The tautomerism of heterocycles. Academic Press, New York
4. Martin YC (2009) J Comput-Aided Mol Des 23:693–704
5. Pauling L (1939) The nature of the chemical bond. Cornell University Press, New York
6. Sayle R, Skillman G (2002) Hooked on protonics. American Chemical Society Fall Meeting Boston 2002. http://www.eyesopen.com/about/events/presentations/acs02/index.htm. Accessed 3 Mar 2010
7. Labute P (2005) J Chem Inf Model 45(2):215–221
8. Lexichem Manual, Version 2.0. OpenEye Scientific Software, Santa Fe, New Mexico. http://www.eyesopen.com/docs/pdf/lexichem.pdf. Accessed 3 Mar 2010
9. Sayle R, Nicholls A (2006) J Comput-Aided Mol Des 20:191–208
10. Taylor P (1998) Tautomeric preference: survey and guidelines. Lecture notes. Los Alamos National Laboratories (LANL), Los Alamos, NM, USA
11. Stocks M, Alcaraz L, Griffen E (2007) On medicinal chemistry. Sigma Aldrich, St Louis

12. Perez GV, Perez AL (2000) J Chem Ed 77(7):910–914
13. Pearlman RS (2005) System and method for providing a canonical structural representation of chemical compounds. United States Patent Application 20050125210(A1)
14. Pearlman RS (2005) System and method for identifying structures of a chemical compound. United States Patent Application 20050159900(A1)
15. IUPAC (1996) Basic terminology of stereochemistry, IUPAC Recommendations. Pure and Appl Chem, 68(12):2193–2222. http://www.iupac.org/publications/pac/1996/pdf/6812x2193.pdf. Accessed 3 Mar 2010
16. Oellien F, Cramer J, Beyer C, Ihlenfeldt W-D, Selzer PM (2006) J Chem Inf Model 46(6):2342–2354
17. Kenny PW, Sadowski J (2004) In: Oprea TI (ed) Chemoinformatics in drug discovery. Wiley-VCH, Weinheim, pp 271–285
18. Milletti F, Storchi L, Sforna G, Cross S, Cruciana G (2009) J Chem Inf Model 49(1):68–75
19. Trepalin SV, Skorenko AV, Balakin KV, Nasonov AF, Lang SA, Ivashchenko AA, Savchuk NP (2003) J Chem Inf Comput Sci 43(3):852–860
20. Ting A, McGuire R, Johnson AP, Green S (2000) J Chem Inf Comput Sci 40(2):347–353
21. Mockus J, Stobaugh RE (1980) J Chem Inf Comput Sci 20(1): 18–22
22. Sayle R, Delany J (1999) In: innovative computational applications: the interface of library design, bioinformatics, structure-based drug design and virtual screening. IIRG publishers, San Franciso. http://www.daylight.com/meetings/emug99/Delany/taut_html/sld001.htm. Accessed 3 Mar 2010
23. QuacPac Manual: Tautomers. Version 1.3.1. OpenEye Scientific Software, Santa Fe, New Mexico. http://www.eyesopen.com/docs/pdf/quacpac.pdf. Accessed 3 Mar 2010
24. Pipeline Pilot Chemistry Collection: Advanced Chemistry User Guide (2008) Accelrys, San Diego, California
25. Moran D, Sukcharoenphon K, Puchta R, Schaefer HF III, Schleyer PR, Hoff CD (2002) J Org Chem 67(25):9061–9069
26. Allen FH (2002) Acta Crystallogr B58:380–388
27. Gilson MK, Gilson HSR, Potter MJ (2003) J Chem Inf Comput Sci 43(6):1982–1997
28. Katritzky AR, Pozharskii AF (2000) Handbook of heterocyclic chemistry. Pergamon Elsevier, Amsterdam
29. Valters RE, Flitsch W (1985) Ring-chain tautomerism. Plenum Press, New York
30. Altshuler DL, Dickson WB, Vance JT, Roberts SP, Dickinson MH (2005) Proc Natl Acad Sci 102(50):18213–18218